**关于Java的面试题：关于虚拟机的面试题：**

1.hashmap底层?

2.Java进程中断怎么处理?

3.介绍下面向对象?

4.介绍下多态?

5.方法重载和重写?

6.这俩哪个属于多态?

7.说下多线程?

8.实现方式有哪些，区别?

9.同步锁 互斥，为什么常量不能做锁?

10.说下反射，都是做什么的?

11.反射机制怎么实现的?

12.linkedlist和arraylist的区别?

13.hashMap的存储机制和存取过程?

14.java的多线程？

15.java的流处理？

16.socket编程？

17.hibernate的原理，干啥的？

18.服务器之间如何通信？

19.定时器如何写？

20. Hashmap hashset treemap 区别

21.什么是hashMap？

22. 常用标签html？

23.Spring  控制反转 依赖注入  ，AOP 及其用途?

24.java 中 怎么解决多线程问题 有几种方式 ，他们之间的不同?

25.Hashtale hashmap currenthashtable 区别?

26.Jvm 机制 以及OOM发生的情况 简述一下?

27.Java 数组反转实现?

28.Java  序列化 SqenceFile 并说明你的序列化方式和适用场景?

29.Hashmap怎么实现，与tremap的区别?

30.Java中问到常问的new String()问题 到底是创建几个对象，StringBuffer和StringBuilder区别?

31.java反射比较多 怎么优化处理?

32. java 集合set  list  map  原理 区别  如何选择？

33.jvm调优？

34.线程同步？

35.web  service生命周期?

36.对spring的理解?

37.java7新特性?

38.java实现上述功能，在TB级别数据情况下，实现方案?

39.java实现取出两个文件中共有的数据?

40.做一个排序（java伪代码）?

41.常用的jqury框架，比如jqury的表单校验框架？

42.单例？

43.LinkdList和ArrayList的区别？

44.http传输的时候是明文的，怎么解决安全问题？

45.垃圾回收机制？

46.java内存模型？

47.java内存模型？

48.java垃圾回收机制?

49.object中有哪些方法?

50.ArrayList中Array的长度超了是怎么增加的，一次增加多少?

51.java多线程?

52.java中的集合?

53.垃圾回收机制?

54.java中的排序算法?

55.会不会SSH框架?

56.线程池怎么实现？

57.Java内存错误中的一个错误，问我如何解决。？

58.Callable和Runable的区别？

59.在线项目的bug如何调试？

60.进程和线程的区别，线程如何配置？?

**关于设计模式的面试题：**

1. 设计模式?

2.说说常用的设计模式，我说了个单例，他好像觉得太简单了，我又说了个策略模式？

**关于数据结构和算法的面试题：**

1.再一次提起数据结构和算法

2.算法和数据结构--要求，是基本的?

3.一个数组，如a=[1,2,3,4,1,2]，把元素出现两次的保留，最后得到a=[1,2]；最后要写几个测试case。

这个问题我先用一个循环，map统计次数，让后再循环。当出现两次的元素，踢掉。

一个字符串，如何查询是否含有某一个子字符串，如果有返回索引，不能用api的函数

4.PageRank怎么实现?

5.算法中O(n)一定比O(lg(n))性能差吗？为什么？有那些因素需要考虑？

算法中O(n)一定比O(lg(n))性能差吗？为什么？有那些因素需要考虑

6.出了个题让我现场搞一下，一个文件里有两个字段分别是IP和time，ip可以通过写好的函数转换成省份，让我实时统计每一分钟的PV，UV，注意PV,UV是累加的，不是一分钟的数据，可以根据省份去查询PV，UV。

7. 对一个字符串进行全排列？

8.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？？

9. 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？

10.一个100G文件，里面每一个行都是一个ip，怎么写单机程序找出出现次数最多的前100个IP？

11.写个二分查找?

12.写程序找出两个字符串中出现的共同字母?

13.快速排序?

**关于HDFS的面试题：------ok**

1.hodoop写文件流程?

2.hdfs了解多少？读写流程，架构？

3.写流程中备份三，其中一个写失败了怎么弄的？

4.hdfs HA（过程，启动流程）？

5.hadoop的优化？

6.hadoop1和hadoop2的区别？

7.hadoop有哪些组件？

8.hadoop 数据倾斜问题?

9.HDFS 存储的什么类型?

10.Hadoop1.x 和2.x 的差别?

11.说说你hadoop 中都每个配置文件都配置了什么吧?

12.Hadoop集群的优化?

13.hdfs的实现?

14.hdfs文件创建--工作流程？

15.hdfs异步读?-------------参考博客：<http://blog.csdn.net/androidlushangderen/article/details/52452215>

16.hdfs api 源码 解答：文件创建工作流程?

17.hdfs  api新特性?---------不知道

18.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？

19.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？-------不知道

20.Zookeeper怎么实现Hadoop的高可用?--------不熟悉

21.Hadoop体系？

22.MR的过程？

**关于MapReduce的面试题：----(ok)**

1.mapreduce的原理和架构？

2.hadoop的全排序和局部排序和二次排序？----需要注意

3.Mapreduce 过程？

4.Mapreduce 优化？

5.通过mapreduce 实现两个表的join?

6.MR的原理 拿wordcount来讲，以及优化(全面些)?

7.Mr中啥时只有map，啥时要有map和reduce?

8.给一个案例  用mapreduce去实现，，这个实现过程，会遇到什么问题，怎么解决？

9.mr数据倾斜解决办法？

10.mapreduce的pagerank实现?

11. mapreduce  merge?

12,用代码写出你使用过的mapreduce？

13.MapReduce流程基本原理？

14.MapReduce过程的数据倾斜怎么解决?

15.MapReduce过程，Zookeeper实现高可用的原理？----不是很清楚

**关于yarn的面试题：-----ok**

1.介绍下yarn?

2.yarn有哪些组件，调度算法？

3.说说yarn工作流程?

4.yarn优缺点?

**关于linux Shell的面试题：**

1.shell熟不熟？文件查找用什么命令？文件内容过滤用什么？grep命名用过没？

2.shell脚本的编写 现场笔试  二个文件  加入都是key value 相同key的value求和?

3.awk的编写?

4.shell 脚本?

sort ， uniq -u，uniq -t ，uniq -d， cat 组合使用，解决从不同文件中找出相同数据的问题，还有解决只在一个文件中出现的数据的问题

5.linux文件中查找字符串的命令。还有替换字符串的命令。还有 awk命令？

6.比iptables更深入的权限控制的啥。。。（说了个我也没记住）？

7.linux里一个文件，怎么替换某个单词的内容，一个文件有几行数据，怎么直接查看第二行数据

8.请使用awk, sed命令对文本文件中第二列和第三列取出来？

9.阿里巴巴的电话面试问到了linux的详细启动过程？

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

10.在linux环境下怎么查看一台机器的配置情况，比如磁盘，内存等

11. Shell编程获取时间，crontab五个星号代表什么？Sed和Awk程序的编写？

12.Shell获取某行第几个字符怎么做?

13.让我写shell脚本求一个文件中的平均值?

**关于hive的面试题：sql语句要清楚**

1:hive的优化?

2.写hive建表语句?

3.简单说下hive？

4.底层实现？

5.hive的表类型？

6.hive创建表的几种方式？

7.udf和udaf都写过哪些?

8.介绍下udaf?自己写过吗？

9.hive用哪个版本?

10.hive窗口函数？

11.写sql：1月100,2月200,3月100,4月200.统计如下效果：1月100,2月300,3月500,4月600.【就是每月统计一次前面所有的月的总额】 加个over，就可以orderby 与partitionby类似？

12.hive 2.0 lllip tz 了解吗？新特性？

13.hive时，怎么合并小文件来着?

14.列出了三张关联的表，其中一张表有点击数量的统计。让我们算一天的所有点击数量写出 hivesql，我没写出来?

15.Hive 的优化?

16.Hive  左右连接?

17.Hive 使用的三种形式?

18.Hive 分区?

19.hive的工作原理?

20.Hive中的rlike是怎么回事，哪里用到了?

21.sql优化

22.问我Hive怎么解决数据倾斜? MapReduce过程?

设计角度  ---  建索引，建视图

sql---------------  充分利用索引，避免全表扫描

22.hive优化?充分利用临时表,重复利用?

23.几道sql问题，联表查，exists，group by保留全部字段?

24.不用hadoop  自己设计方案  实现TB级别数据量，TOP10问题，数据倾斜问题怎么解决?

25.hive原理，优化？

26.hive数据倾斜优化？

27.hive 执行sql时的工作流程?

28.hive的默认分隔符?

29.hive rank（以某个字段分组，在组内排序，找出各组的TOP  k）?

30.问到了hive的优化？

31.怎么解决HIVE产生的数据倾斜问题?

32.问我Hive具体怎么用的?

33.HIVE中UDF UDAF UDTF的区别。数据倾斜问题怎么解决？表连接有几种类型？

34.HIVE怎么优化？HIVE常用的几个配置是哪些？

35.Hive什么时候不能用combiner？

36.添加自定义UDF的时候都是临时的，怎么永久添加？

37.Hive动态分区?

38.写一个表的查询的sql语句，具体忘了，是个嵌套的SQL?

**关于Hbase的面试题：------(ok)**

1.hbase 我说没用过?

2.hbase有哪些组件，以及hbase里面的buffer？

3.hbase rowkey的设计？

4.hbase优化？

5.hbase数据插入优化？

6.hbase如何优化的？

7.hbase的writebuffer？

8.hbase对于复杂表如何设计rowkey?

9.Hbase 执行过程？

10.Hbase 优化？

11.关系型数据库是怎么把数据导出到Hbase 里的?

12.hbase 优化问题?

13.hbase 的性能优化?

14.Hbase解决了什么问题?

15.Hbase的相对多些，基础和优化?

16.hbase最主要的特点是什么？

17.hbase部署?

18.和hbase同样功能的分布式数据库了解多少?

19.HBase的存储原理?

20.说说hbase的API都有哪些filter?

21.问我major compack设置为手动会出现什么问题？

22.mongoDB和hbase的区别？

23.Hbase的表设计还有优化?

**关于Storm的面试题：------------(ok)**

1.说说storm？

就讲Storm的各个方面：

Storm是什么？架构中为什么要选择它而不选择SparkStreaming？Storm的编程模型？Storm的架构

设计和Storm的容错机制？

我在项目中是怎么使用Storm的？

2.strom窗口：五分钟统计一次。？

3.storm原理?storm如何保证不丢失数据?

Storm的编程模型:Topology(DAG有向无环图的实现)：spolt，bolt，构建topology，storm的并发机制及数据分发策略

Storm保证数据不丢失是Storm的有保证消息的完整(tuple树)处理的机制：acker机制(ack的实现原理:通过tuple的id的亦或运算来判断消息是否被完整计算实现,所以在spolt发送tuple的时候需要设置消息的id)，但是这样会导致消息的重复计算，storm提供了拓扑性的事务(分阶段来实现事务的强有序和并发性)来保证消息有且仅被处理一次

4.storm 流处理数据丢失？

一般不会丢失，Storm大多的bolt都实现了acker机制，保证数据不会被丢失，当数据丢失的时候，acker机制会回调ack方法和fail方法重发tuple。

5.Storm的原理?---同3

编程模型

6.spark streaming和storm的区别，可以相互取代吗？

纯流式的实时的计算框架和微批处理的框架

spark家族一栈式的大数据处理框架，storm显得很专业

事务支持方面：Storm事务支持的比较好，SparkStreaming差点

实现的功能方面：SparkStreaming提供丰富的算子可以实现丰富的功能，Storm一般做比较简单的统计

7.说说你用过的storm?---同1

8.storm的设计和日志的格式？

Storm的设计主要是对pv，uv等简单的统计的topology的构建，还有其并发的设置

日志的格式：

9:storm整合kafka？

Storm的Spout应该是源源不断的取数据，不能间断。那么，很显然，消息队列系统、分布式内存系统或内存数据库是作为其数据源的很好的选择

由于storm-kafka已经实现了spout，我们直接用就可以

**关于Zookeeper的面试题：**

1.问了Zookeeper的工作原理。过半机制，还有节点为什么是单数台？

2.zookeeper熟悉不？

3.zookeeper有哪些组件？

4.zookeeper的机制等，各组件的原理？

5.Zookeeper到底是干嘛的?

6.zookeeper的二次开发？

7.Zookeeper实现分布式锁用哪个jar包，以及写mr、spark作业程序具体应该用哪些包？

**关于Kafka的面试题：------**

1.怎么保证kafka传过来的数据之正确的处理一次？-----结合Storm事务来思考

2.flume 和 kafka什么区别？

3.kafka为什么要分多个partition？

4.kafka和spark Streaming 的整合？-------重要-----不是很清楚，看kafka和SparkStreaming整合

5.怎么保证数据kafka里的数据安全？（丢失）----磁盘存储，数据使用完后的删除的策略

6.kafka的key为null可以吗？

7.怎么往kafka集群写数据的？------Kafka Sink(使用的是Flume)或者KafKa Spout(如果使用的是Storm)

8.kafka用到的什么设计模式？----发布订阅模式

9.kafka的原理?如果生产数据是消费数据100倍，该如何处理?

10.flume与kafka区别?

11.有很多消息队列技术，为什么选择kafka ?----kafka的特性方面回答

12.kafka为什么可以支持那么大的吞吐量，怎么实现的，我直接说不知道。?----顺序读写，partition的分布式存储

**关于Flume的面试题：-------ok**

1.flume什么时候用？----分布式的数据收集

2.Flume 处理日志重复问题?

3.flume 监控问题  ?自身支持ganlia集群监控

4.Flume是如何拿数据的?----flume source

5. flum是如何导入数据到kafka?具体?Kafka Sink channnel里面配置相应的topic

6.flume的实时采集数据和定时采集数据的方法？tail -f；Spooling

7.主要问Flume如何使用？整合Flume+Storm和整合Flume+Kafka

**关于Sqoop的面试题：-----ok**

1.sqoop导数据 是否了解sqoop？了解，通过sqoop export/import来实现

2.sqoop底层（我说不参与）？了解，通过MR来实现的数据的导入导出

**关于Spark的面试题：------ok**

1.spark优化?

答案：

架构参数优化：shuffle，内存管理，推测执行，数据本地化：HDFS的DataNode和Spark Worker共享一台机器

代码层面：并行度--调整finalRDD partition；缓存机制的选择--CPU使用和内存使用的权衡： checkpoint；算子的使用和选择-groupbykey，map vs mappartitions等，使用广播变量，累加器等； 序列化：压缩，存储格式的选择

数据倾斜：重写partition规则，抽样看数据的分布，结合具体的业务

架构的选择：统一使用yarn结合hadoop，还是使用自己的standalone计算框架

1.spark的工作流程？

答案：Spark的资源调度和任务调度+pipeline的计算模型

2.spark源码-DAG-Task--任务调度部分？

答案：

首先清楚spark是粗粒度的资源申请，任务调度：sparkContext-DAGSheduler切分stage，TaskSheduler发送任务到申请好的Executor中的线程池执行

3.submit相关配置？**一般指定多大的资源？**

答案：submit --master/yarn --class --deploy model clster/client

--Executor cores 默认一个Executor 1 core，lg内存，1G，2--3个task

4.写完spark程序如何知道多少个task？ （即资源如何调配的）

答案：看你的并行度的设置，block的数量，web UI

5.spark和mr性能是不是差别很多？

答案：

一般来说Spark比Hadoop快：

原因：

（1）：MR有大量的磁盘io,溢写等，Spark则可以基于内存缓存机制计算

（2）：MR和Spark的资源申请的方式：粗粒度和细粒度的区别

（3）：DAG计算引擎中的pipeline计算模型，MR就是MapReduce模型

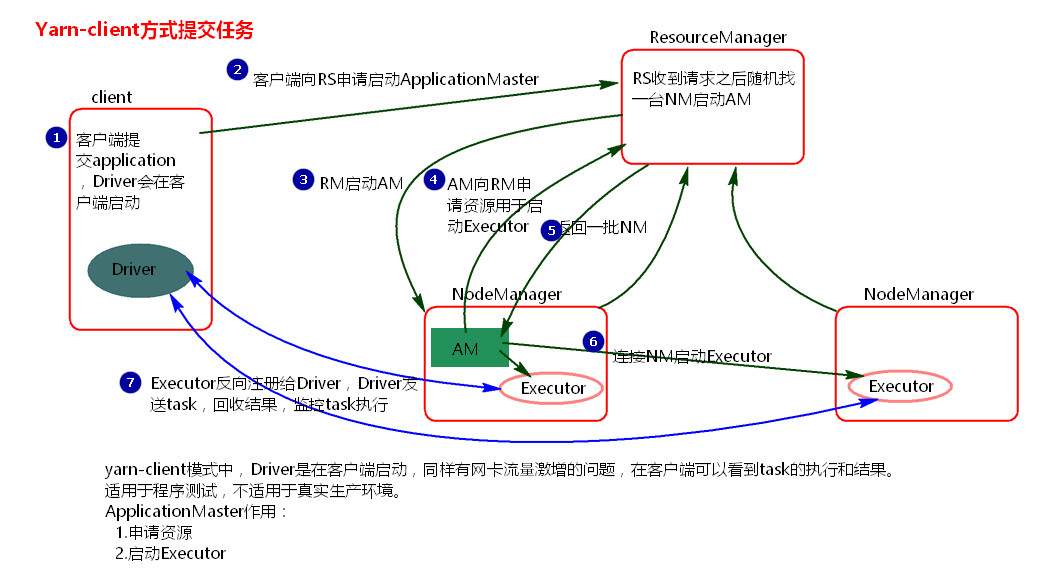
（4）：算子的丰富程度

使用场景：大于pb级别的数据量一般选择MR

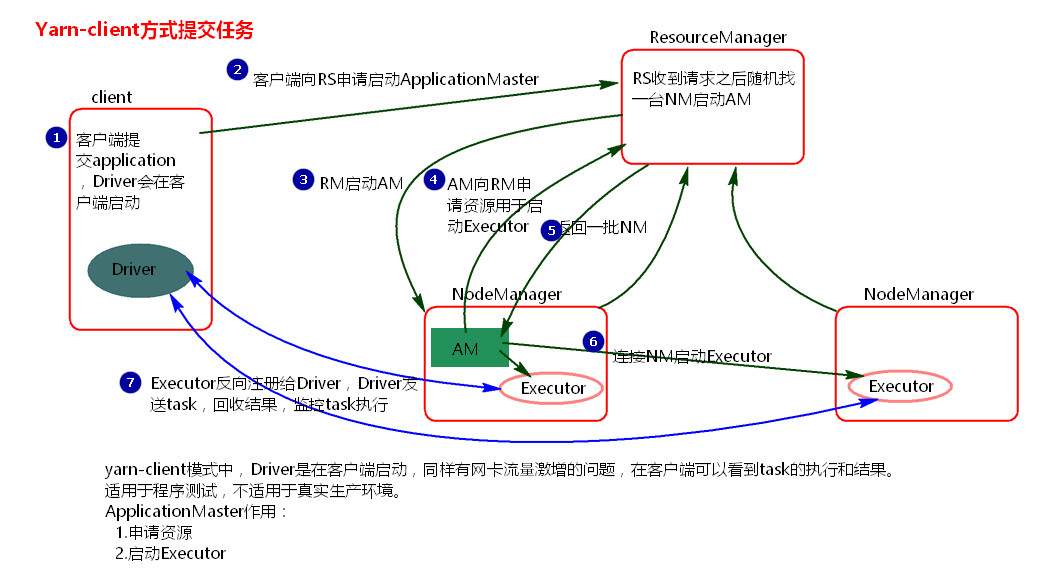
生态的区别：Spark一站式的大数据处理平台，Hadoop还需要和其他的整合，升级，版本兼容等一堆问题，CDH版本如果需要更多的功能需要考虑成本的问题

6.spark任务yarn执行流程(client)?

答案：



7.spark运行在Yarn上流程（cluster）?



使用场景的区别：基于yarn的好处，兼容hadoop，一套计算框架，能好的维护

8.spark调优？----同问题1

9.shuffle主要介绍下？

答案：shuffle发生？---shuffle的过程---shuffle实现的选择---shuffle的优化

10.宽窄依赖？

答案：看父RDD和子RR的关系，除了父RDD和子RDD一对多外，其他的都是窄依赖

11.shuffle怎么落地的？

答案：shuffle的实现类型：hash Shuffle还是sortShuffle？Shuffle数据落地？

12.Spark RDD 是什么?

答案：弹性分布式数据集---源码的五大特性-----RDD的计算模型：pipeline计算模型

13.Spark 算子?

答案：map，flatmap，filter，foreach，first，take(n),join,cogroup,reducebykey,sortBy,

distinct,mapPartition等等

14.spark 优势？

答案：

一栈式大数据处理平台。

灵活的编程模型，相比MR

速度快

15.spark on yarn  和mapreduce  中yarn有什么区别？

答案：没什么区别，yarn就是一个资源管理框架

16.spark 原理？

答案：pipeline计算模型+任务调度和资源调度

17.spark运行的job在哪里可以看到？

答案：Driver进程所在的节点；web UI

18.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?

答案：Spark WEB UI；集群监控工具，找到taskid

19.spark为什么比hadoop快？同题5

20.rdd的处理过程是什么，不要说概念?

答案：画切分Stage，pipeline的计算模型的图

21.请说出你在spark中的优化方案？

答案：同1

22.SparkSQL和Spark架构，运行流程图，Spark运行的两种方式。常用的Spark函数有哪些？

答案：spark架构图+运行流程图(资源的调度+任务调度)+Spark client和SparkCluster+transformation算子+action算子+持久化操作算子

24：GroupByKey的作用？

答案：根据key分组

23.Spark了解多少？

答案：Spark生态-架构-运行模式+任务调度和资源调度

**关于Spark Sql的面试题：**

1.sparkSQL介绍下（RDD、DataFrame）

**关于Spark Streaming的面试题：**

1.sparkStreaming怎么跟kafka对接的,数据拉取到哪里？

2.日流量10G没必要sparkstreaming？

3.spark streaming 例子。问维护做过没？说sparkStreaming的维护成本很高。 我告诉他是的，比如说可能会丢数据，wal会慢。这一块儿不是我维护。没细问。

4.spark streming调优?

5.sparkstreaming原理？

6.spark Streaming介绍下？和Storm比较？

7.spark Streaming某一个task挂了，怎么解决的?

8.spark streaming?spark的相关算法，比如推荐系统需要什么算法?

9.park streaming 工作流程？

10.sparkstreanming没有问题，但无法计算，怎么排查问题?

11.storm和spark streaming的区别？

**关于Spark MLlib的面试题：**

1.spark MLlib那部分也问了我很多，因为他没搞过机器学习，所以这部分回答的问题不大。

**关于机器学习的面试题：**

1.机器学习的数据量级别?

2.Adaboost 算法,讲讲你懂的算法？

3.让你写一个机器学习的项目能自己写出来吗？

4.机器学习的算法?

5.机器学习项目用什么写的？

6.机器学习各种算法都了解吗？接下来问的聚类算法，k-means

7.机器学习是不是不能用mr？

8.机器学习算法？

9.k-means 的聚类实现?

10.机器学习是怎么回事?

11.k-means算法如何实现，为何收敛？

12.说说掌握那些算法，如决策树，神经网络，知道那些聚类算法?

13.你在项目中做机器学习的时候，遇到的最大难点是什么？

14.问我机器学习了解多少?

**关于scala的面试题：-----ok**

1.用scala写一个wordcount？

答：textfile().flatmap(line.splite()).map(key.1).reduceByKey(v1+v2)

2.对scala的了解 有用过play framework 框架吗

答：熟悉scala，编写过一些spark应用程序，就是使用scala编写的，还有看spark源码的折腾出来的

我知道它是scala 的一个web端的开发框架，好像还有一个叫Lift，但是我没用过，不是很了解

3.写scala程序，主要是处理输入文本方面。过滤特定数据，按照指定顺序输出?

答：textfile().filter().sortBy().foreach()

4.scala的变量？他说函数编程一般没有变量，scala是变种啥的？

答:scala的val常量和var变量，。。。。。

5.scala的闭包？

答：闭包是一个函数，返回值依赖于声明在函数外部的一个或多个变量

6.scala中的隐式函数的关键字？

答：implicit

7.val x=y=1结果是什么？

答：会报错，如果要多个变量同时赋值：val x,y=1

8.编译好的scala程序，运行的时候还需要scala环境吗?

答：不需要，scala编译成.class文件，运行的时候只需要jre环境即可

9..Scala一些基础的问题，如：伴生对象，类的问题，有哪些class？

答：同一个文件中：object和class的名相同，class叫object的半身类，object叫class的半生对象

样例类：在模式匹配的时候常常看到样例类。

**关于Redis面试题：**

1.redis用来做什么？ 模型等，频繁调用的放在redis中，取其快

2.redis的常用数据类型？

3.redis 工作原理？

4.redis缓存机制?

5.redis支持的最大数据量是多少？redis集群下怎么从某一台集群查key-value。

**关于项目的面试题：**

大数据平台：

实时推荐系统：

1.项目流程,机器学习的项目流程,电商项目的数据流程？

2.你们一个work给分配多少资源？怎么分配的,预先分配吗?

3.怎么收集的数据？

4.你项目都负责哪一块？

5.推荐系统建模周期,这期间遇到过什么问题？

6.sample正负例样本表，标签是怎么打的？

7.数据来源是什么？

8.标签值是不是不多？（正负例样本表是标签+-1），他指的标签是维度

9.期望薪资?

10.讲项目？

11.项目数据量，机器学习的项目肯定不大？

12.模型auc直多大0.92，他说挺大，我说我调的准，混淆矩阵相关算法，怎么算的？

13.还有服务器多少台?

14.介绍最近的项目？

a.什么是协同过滤，

b.协同过滤的值怎么求得，

c.hive的调优，

d.具体的pv/uv量，

e.训练数据量，

f.有多少个维度，

g.特征怎么选取的，

h.模型效果怎么评估）

15.另一个项目问到了数据怎么收集的？

16.埋点怎么弄？

17.你具体负责哪一块？

18.剩下的俩项目你选个讲吧？

19.推荐系统那一套？负例少，正例多怎么办？

20.对自己每个项目做讲解，项目中的疑难点？

21.服务器如何选择？项目服务器多少台？namenode多少台？dotanode多少台？kafka多少台？yarn多少台？

22.讲解自己的项目，遇到的问题？

23.问我数据量多大问题，和mapreduce运行时间问题，由于我实现没有准备好，回答不好，订单的我回答50G，微博我回答1TB，mapreduce运行时间我回答 1~2小时？

24.的推荐系统矩阵列表是怎么实现的?

25.你日志处理具体怎么写的mapreduce  流程?

26.storm 项目中遇到了那些问题，怎么解决？

27.用到hbase的项目提问，实际如何处理的，java是怎么调用的，数据太多怎么优化，你所设定的数据要处理多久?

28.如何搭建实时日志分析平台，需要那些条件？

29.从设计架构，业务实现，为什么这样做，性能如何，等等问题，很多地方深入到项目中实现细节？

30.训练集和测试集的比例多大？

31.描述一下逻辑回归的特点?

32.说说项目中用到的框架？

33.项目里的业务啥的谈了谈？

34.两个项目电信和交通厅，分别用了什么架构，怎么搞得，参与搭建了吗？

35.接着又问flume几台，怎么从其他系统获取的数据，kafka几台？

我说的kafka吞吐量10万条信息每秒，我们用了一台，接着问那一台kafka挂了呢？

36.这个地方回答的不好，没搞过kafka高可用，说多台kafka也是坑，到处都是陷阱。

37.项目中那块是你做的？我说的storm实时通话分析那里，问storm怎么从kafka里读数据的

接着又问storm的spout几台，我说一个，接着说spout挂了怎么办？实在没法回答这些破问题，根本都没遇到过，吹的话那继续深入的问，一堆坑。

38.storm处理完数据写入哪里？回答hdfs和hbase   又问storm怎么写入hdfs和hbase的具体说一下。

39.问我交通厅项目主要做了哪些部分？我说spark MLlib预测路况那部分，问用的什么算法，我说逻辑线性回归

40.接着问线性回归的原理，什么场景适合线性回归，举两个例子说下。

41.模型生成完以后是怎么知道预测的好坏的？

42.对了还问了storm处理的时候利用率怎么样？怎么检测storm没有问题的，程序跑通就一定没有问题吗？反正我也不知道怎么回答了，不知道大数据有没有测试人员，怎么测试改需求?

43.自我介绍，然后就项目，电信项目我主要做了那一块？我说strom实时通话分析那块?

44.怎么从其他系统获取的数据，回答flume+kafka+storm这样的流程。

45.接着问flume有几台，通过什么协议获取的数据，然后就开始开火了?

46.flume收集信息的时候遇到了什么问题？怎么解决的？

47.kafka几台，我回答一台，因为kafka最大支持吞吐量10万条每秒，接着问你们kafka传输的实际吞吐量是多少条每秒，一直追问这个，我没遇到过真不知道怎么回答，kafka传输数据的时候遇到什么错误吗？怎么解决的？又是坑，说没有遇到过。接着又问你们kafka处理的时候都没遇到过什么问题吗？弄得我无言以对，沉默。

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

49.你在项目中使用的技术，解决了什么问题？

50.在你做的项目中所使用到得技术或者工具，都是做什么的？

51.flume在实际项目里面的数据采集？

52.感觉自己工作里面做的最好的是哪一块？

53：关于集群数据量，运行时间的参考

**刚才面试面试官问了你们每天有多少数据，**

用了多少台机器，

一般根据你写的项目，每天产生的数据量规划，假如一天数据量100g

一般集群 规划是年数据的3倍，还有 hadoop集群3倍冗余

假如一台服务器磁盘6T

100G\*365\*3\*3/6\*1024g=53.4  这样的集群（一般在60台左右的服务器）

机器的配置，

配置 cpu 找一个稍微老一点至强cpu

内存每台16g或32g

**每天运行多久**

一般一个作业10分钟到-几个小时不等

一般一个作业也就几十分钟。。运行几天的很少

**有多少个MR，**

30-50个左右

一般公司很多个作业。。

你可以你们部门的,其他你不清楚就别说，比如数据清洗的（这里面就有很多作业了，去掉不完整数据，数据格式转换，数据字段连接，字段抽取等等），相应你简历上写的项目，，很多模板都有作业。。你细化一下

比如推荐的作业，统计汇总的作业，用户定位的作业，

**遇到bug怎么解决，上线之后的bug怎么解决，**

一般在测试阶段就那部分线上数据测试过了。。

如果在线上还有问题

一般kill掉作业。。当然可以做mapreduce里面设计日志输出到单独文件，，

根据hadoop异常日志出什么问题了。。当然hadoop每台都会有日志，当然hadoop自己的日子很庞大，，可以采用chukwa（大概看看干什么的就行，就是收集方便查看hadoop本身的日志）处理，

然后分析作业代码。。

**有没有关心过运行时候的状态，**

mapreduce运行状态，，hadoop有监控页面

当然也可以自己写监控程序，，mapreduce有作业监听方法，可以获取进度。。

<http://m.blog.csdn.net/blog/u014313009/38045435>

**其他面试题：**

1.各个软件的版本号？

2.spark程序用什么语言写的？

3.用Python写的什么？

4.为什么辞职?

5.最快什么时候可以入职?

6.标签数据库?

7.为啥来北京上班。。。。？

8.数据倾斜在什么时候发生？

9.数据清洗流程（源，过程）？

10.nginx了解吗？（没搞过，不负责）？

11.数据清洗怎么发现的？

12.清洗完后面都有哪些要求?

13.DSL和SQL用哪个比较多?

14.为什么会发生数据倾斜，你怎么知道发生数据倾斜的？

15.事物方面比较详细说下？

16.介绍自己，讲讲自己的项目？

17.hadoop的生态圈？

18.会配置本地yum源码？

19.是否用过maven？

20.谈谈redis？

21.平实喜欢关注什么关键技术，论坛?

22.用到哪些全文检索的技术?

23.免密码登陆如何设置?

24.linux命令，都是平时用到的命令?

25.是否配过yum本地源?

26.是否熟悉python，scala等等？

27.lunce 和solr？

28.你的集群多大 每天流量多少？

29.你集群中定时任务是怎么做的？

30.监控系统做过没？Ganglia 监控系统 或者集群？

31.用python 写一个脚本程序 向mysql 批量添加用户?

32.Mysql 主从复制一段时间后 突然数据同步不了怎么办?

33.说说你的都用了naginx 中都用了那些模块在什么条件下使用的?

34.Lvs解决什么 然后就是画图说负载的问题  mapreduce的执行过程  有木有带过团队?

35.lvs和nginx比较？

  浮动ip  ，负载均衡

  面试官说  ，实际nginx就足够了 ，不需要lvs做高可用

36.什么不用hive呢？

37.ha原理？

38.远程通信技术？

39.对于宜信宜人贷平台的认识?

40.给与用户的每天通话数据，如何在五天内给出一个可行能够评判用户信任度的系统

主要是针对具体业务的技术解决方案，可靠性，备用方案等?

 我给了两种方案，人工标注部分数据做训练集，做分类算法，被否；然后给出聚类算法，从用户特征到聚类实现，可靠度评测等，面试官认为结果不准确，可以接入用户信息，并让我给出解决方案，这个讨论一半，面试官去开会，加了微信说是之后再谈，因为是猎头推荐的，后续猎头说继续跟进，暂时没有反馈

41.eclipse版本号?

42.大数据部门，说是让去独立负责这一块?

43.项目业务场景，问题，架构，人员分配，项目中遇到问题，规模等等----

再然后问了许多  为什么辞职，而且一直在追问具体情形（我回答是：希望换个环境，最后逼问没办法，只得回答说是，期望有一个更好的发展，然后说下自己的规划什么的）?

44.聊聊SOA，实现技术，最基本的特征啥的？

45.说说kettle的简单流程？

46.谈谈lvs和keepalived，怎么用的？跟其他负载均衡的区别？有啥缺点？具体用在什么情景？

知不知道另外一种负载均衡的技术，比lvs要好，能同时作用与第四层和第七层。还说lvs一般只做DB的负载均衡

47.oracle的优化，可以自主选择索引的那种？

48.面试官问我以前的企业有多少人怎么说？问组内有多少人怎么说?问以前项目的数据量怎么说？问集群节点数量怎么说等等一系列比较菜鸟的问题，这些问题说重要倒不很重要，但是说不重要的话如果人家突然问到，你胡乱回答却容易闹笑话，让对方瞬间揭穿你毫无经验的事实。所以我的回答一般都是说用的测试数据，项目上线时候我本人没有参与。也许这么回答不是那么精妙，但是站我自己的角度感到最起码没有大的问题。（比如笔者第一家面试的时候，人家问我当初写的mapreduce程序运行一次多久，我说2小时。。。然后对方表示很震惊，问我多大数据量，我顿时碉堡了，感觉说多少都不合适，于是就找个机会岔开了一下话题，我说您是想问我mapreduce调优这一块吧？不等对方回答，我赶紧把背好mapreduce调优说了一下，来回胡扯了几句，结果竟然混过去了对方最后没再追问，还给了offer。。。第2家开始，所有面试问到运行了几分钟时候我一般会回答1到3分钟之间，项目用的测试数据，所以数据不大，节点的话就依照平时咱虚拟机装的节点规模稍微大一点

49.问了下openstack，我回答了下openstack的各个模块和作用

50.hbase+solr，让我说了一下，解释了一下，问solr你们怎么建索引的，建了多少索引，根据什么去建的这些索引？最后还问solr的索引表是单张表，还是多张表，这些表是存在hbase里面了还是分开的？为什么没有存hbase里面，搞得我无言以对。

51.自己熟悉大数据的部分说一下？

52. hadoop与storm、spark的比较？?

53.事务都有哪些特点？

54.自我介绍？

55.问到了还使用过其他什么开源框架.？

56。问到了最擅长那种技术？

57.问到了在实际开发的过程中遇到了什么问题？

58.人事经理问了许多关于工资的问题,但是由于没有真正工作过,有很多地方说的不合理,之后才了解.有通知,但是一些工作的常识问题没有答上来,就黄了？

59.会不会python?

60.平时遇到问题怎么解决?

61.非技术性：是否愿意出差，平时爱好。。。

62.是否参加过CodeReview？有什么心得？

63.Hadoop用的什么版本？他们公司用的是商业版?

64.PV UV IP解释?

65.数据倾斜问题怎么解决?

66.就问我会不会写PLSQL，好像只是在招写SQL的人?

**67.项目经验**

这一个项目是什么，这个项目你负责了什么

你解决什么难题