mapreduce流程基本原理：

<http://blog.jobbole.com/84089/>

<http://www.cnblogs.com/zhangchaoyang/articles/2648815.html>

<http://www.cnblogs.com/dyllove98/archive/2013/06/08/3127526.html>

<http://blog.csdn.net/wisgood/article/details/16897351>

**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>

<http://www.aboutyun.com/thread-9781-1-1.html>

<http://www.tuicool.com/articles/uM3Iv2>

<http://www.bkjia.com/yjs/962189.html>

用户可以写程序通过jmx方式获取hadoop 监控数据。。然后做相应逻辑处理（比如作业卡死了，执行不成功，可以设计发邮件，发短信程序，当然这个都是程序员自己做的。）。。下面这篇文章一定要看。。。

<http://slaytanic.blog.51cto.com/2057708/1179108/>

**每台机器的负载之类的**

采用ganglia,nagios,zabbix监控工具监控机器磁盘，内存，cpu..你只需要回答采用这些弄得

具体运维部弄得。。当然你研究过会更好

--------------------------------------

  spark sql的任务如何前台有查询请求时提交任务，学的时候一直是local模式或者shell输入命令  
 设计一个作业啊。。  
把sql当做参数穿进去啊  
spark提交作业的时候。。最后一个不就是参数吗？  
程序员，不随你怎么设计吗？  
 local 模式只是本地模拟测试。。去掉local  
导出jar，就可以提交到集群运行了

spark作业远程提交。

两种方式

1.把jar打包好，放到相应spark作业提交服务器，通过java程序远程执行 linux提交。。

2.采用代码执行远程提交，指定对master,相应\*-site.xml放到代码src下面。。也可以提交。参考下面两篇文章

<http://my.oschina.net/132722/blog/220096>

<http://my.oschina.net/132722/blog/196027>