1. 讲一下 solr 吧

嗯,solr 在我们项目中主要是做站内搜索的时候用的 Solr,我知道他底层是基于 Lucene 开发的一个站内搜索引擎，我们项目中一般用它做全文检索,他搜索的时候是通过词条进行搜索,因为 mysql 数据库具有 I/O 读写瓶颈问题，分布式系统的数据量一般比较大，我记得 mysql 单表最大存储数据量是 350W 多条(强调：这个没测试过，是看资料提到过)，一旦数据量过大，必然会引起查询速度慢，服务器资源耗费多等各种问题，所以这时近实时搜索引擎 solr 就是专门解决这种问题的技术实现。

2. 为什么要用 solr

因为 mysql 数据库具有 I/O 读写瓶颈问题，分布式系统的数据量一般比较大，但是 mysql 单表最大存储数据量 350W 多条(强调：这个没测试过，是看资料提到过)，一旦数据量过大，必然会引起查询速度慢，服务器资源耗费多等各种问题，所以这时近实时搜索引擎 solr 就是专门解决这种问题的技术实现。

当然除了 solr 还有 lucence,还有 Elasticsearch， solr 前身就是 lucence,对于 elasticsearch 也是基于 lucence,它的特点的话是实时搜索比较高效。然后 elasticsearch 只支持 json 形式，而 solr 支持多种数据类型格式。

3. Solr 在项目中怎么用的

参考项目话术

4. Solr 分词器有什么效果?

Solr 在搜索的时候本来就是利用词条为索引库进行搜索的,所以我理解的分词就像是索引一样在 solr 索引库中存在,还有 solr 是老外开发的,所以分词是根据英语单词进行自动分词，但是中文的话无法识别，每个字都认为是一个词汇，所以我们在使用 solr 的时候需要配置一个 IK 中文分词器使用的..

5. 关于 IK 分词器扩展词汇实现自动添加用户搜索热词的方法.你有什么见解?

所谓热词就是用户经常搜索到的词语,我们给他放到自己的分词器里当成一个词进行搜索,比如说经常说的”大吉大利今晚吃鸡”,这样的词就可以当做是热词,我也没在项目里实际的做过,但是我知道在大数据里有一个 wordcount 技术可以实现这个热词分析,他是结合的

redis 一起使用的,redis 作为计数器,HBASE 作为数据的存储,然后统计出搜索次数比较多的词条配置到 solr 权重中.

6. 你们 solr 有专门的地方维护吗?是手动添加还是?(同步问题)

我们用的是 SpringDataSolr 来操作的 Solr 索引库.

我们项目在第一次上线的时候,会手动往 solr 索引库导入一批数据,后期就不用人工干预了,我们后台添加了相应的商品之后,商品审核通过的时候，用的 ActiveMQ 往里面发送一条消息,商品的 ID,然后在 solr 这个 search 这个工程中，我们会接收到这个 id，然后根据这个 id 从数据库里面查询出来该商品信息,把数据添加到我们的索引库里边去,增量更新,维护大概就是这样,

7. 全文检索谁来定？

由业务来定。比如说做商品检索时，时搜索商品的名称，卖点，描述。以这些业务域进行所搜的

8. ElasticSearch 和 Solr 的区别?

我知道他俩的底层都是基于 lucene 实现的,都是使用的 lucene 的倒排索引实现的,solr 在实时建立索引的时候会产生 IO 阻塞查询性能会比 ElasticSearch 差一些,还有就是因为 Solr 自身不支持分布式,ElasticSearch 是实时处理数据,而且默认的支持分布式的,可以组成一个网络,如果其中一台服务器宕机,会分配其他节点工作,可以扩展多台服务器,所以查询效率会更快,据说可以处理 PB 以上级别的数据.