金三银四Java面试突击专题

搜索引擎篇

=== 图灵： 楼兰 ===

一、什么是倒排索引？有什么好处？

索引： 从ID到内容。

倒排索引： 从内容到ID。好处： 比较适合做关键字检索。 可以控制数据的总量。 提高查询效率。

搜索引擎为什么比MySQL查询快？ lucence

文章 -》 term ->排序 term dictionary -> term index -》 Posting List -> [文章ID，[在文章中出现的偏移量]，权重]TFIDF

二、ES了解多少？说说你们公司的ES集群架构。

ES： 是一个基于Lucene框架的搜索引擎产品。you know for search。提供了Restful风格的操作接口。 ELK

Lucene：是一个非常高效的全文检索引擎框架。java jar

ES的一些核心概念：

1、索引 index： 关系型数据库中的 table

2、文档 document： row

3、字段 field text\keyword\byte： 列

4、映射Mapping： Schema。

5、查询方式 DSL ： SQL ES的新版本也支持SQL

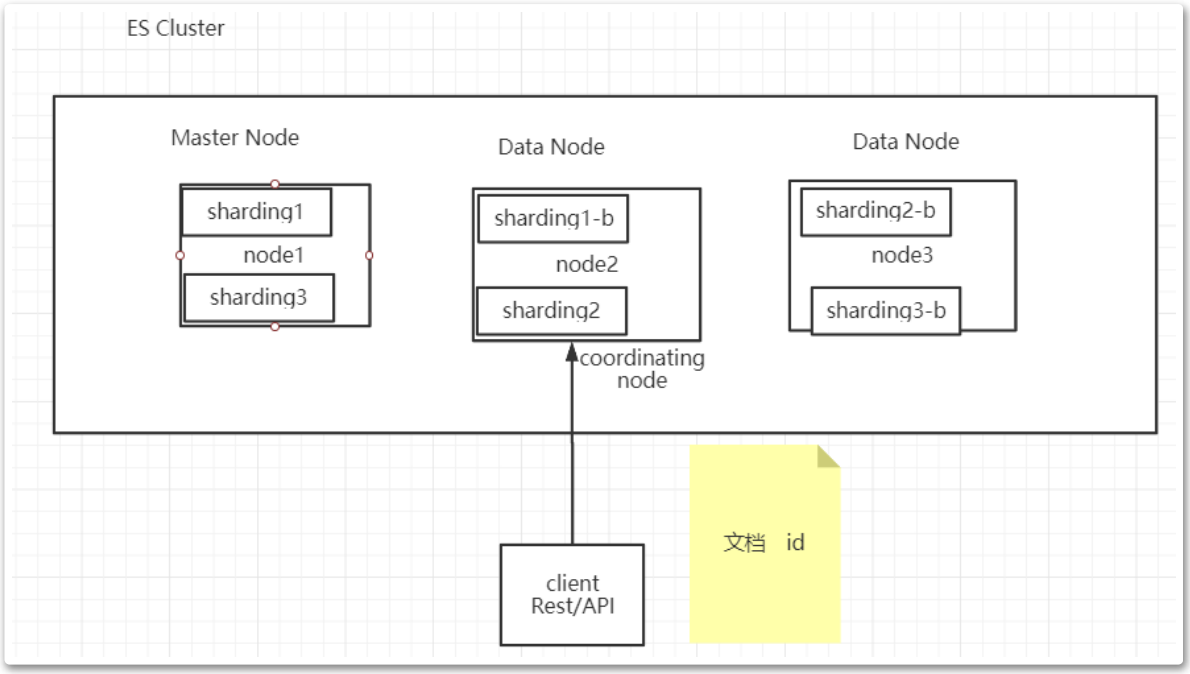
6、分片 sharding 和 副本 replicas： index都是由sharding组成的。每个sharding都有一个或多个备份。 ES集群健康状态：

ES的使用场景。ES可以用在大数据量的搜索场景下，另外ES也有很强大的计算能力。用户画像

三、如何进行中文分词？用过哪些分词器？

IK分词器。

四、ES写入数据的工作原理是什么？



1、客户端发写数据的请求时，可以发往任意节点。这个节点就会成为coordinating node协调节点。

2、计算的点文档要写入的分片：计算时就采用hash取模的方式来计算。

3、协调节点就会进行路由，将请求转发给对应的primary sharding所在的datanode。

4、datanode节点上的primary sharding处理请求，写入数据到索引库，并且将数据同步到对应的replica sharding

5、等primary sharding 和 replica sharding都保存好文档了之后，返回客户端响应。

五、ES查询数据的工作原理是什么?

1、客户端发请求可发给任意节点，这个节点就成为协调节点

2、协调节点将查询请求广播到每一个数据节点，这些数据节点的分片就会处理改查询请求。

3、每个分片进行数据查询，将符合条件的数据放在一个队列当中，并将这些数据的文档ID、节点信息、分片信息都返回给协调节点。

4、由协调节点将所有的结果进行汇总，并排序。

5、协调节点向包含这些文档ID的分片发送get请求，对应的分片将文档数据返回给协调节点，最后协调节点将数据整合返回给客户端。

六、ES部署时，要如何进行优化?

1、集群部署优化。

调整ES的一些重要参数。path.data目录尽量使用SSD。定时JVM堆内存大小。

关于ES的参数，大部分情况下是不需要调优的，如果有性能问题，最好的办法是安排更合理的sharding布局并且增加节点数量。

2、更合理的sharding布局：

让sharding和对应的replica sharding尽量在同一个机房。

3、Linux服务器上的一些优化策略：

不要用root用户；修改虚拟内存大小；修改普通用户可以创建的最大线程数。

ES生态： ELK日志收集解决方案- filebeat(读log日志)-> logstash -> ElasticSearch -> kibana、Grafana、自研的报表平台