1. 简介

Elasticsearch，分布式，高性能，高可用，可伸缩的搜索和分析系统

每个es节点都封装了lucene

<什么是Elasticsearch.png>

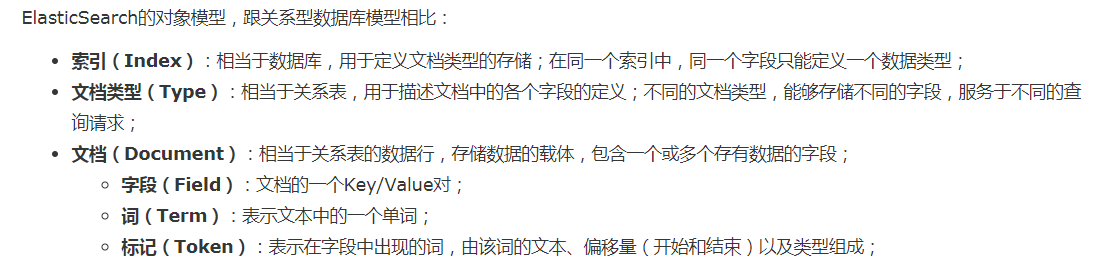
1. 启动

<http://localhost:9200/?pretty来验证是否启动成功>

<http://localhost:5601>

使用kibana来开发

1. 索引/文档



1. 简单的集群管理

es提供了一套api，叫做cat api，可以查看es中各种各样的数据

GET /\_cat/health?v

如何快速了解集群的健康状况？green、yellow、red？

green：每个索引的primary shard和replica shard都是active状态的

yellow：每个索引的primary shard都是active状态的，但是部分replica shard不是active状态，处于不可用的状态

red：不是所有索引的primary shard都是active状态的，部分索引有数据丢失了

为什么现在会处于一个yellow状态？

我们现在就一个笔记本电脑，就启动了一个es进程，相当于就只有一个node。现在es中有一个index，就是kibana自己内置建立的index。由于默认的配置是给每个index分配5个primary shard和5个replica shard，而且primary shard和replica shard不能在同一台机器上（为了容错）。现在kibana自己建立的index是1个primary shard和1个replica shard。当前就一个node，所以只有1个primary shard被分配了和启动了，但是一个replica shard没有第二台机器去启动。

做一个小实验：此时只要启动第二个es进程，就会在es集群中有2个node，然后那1个replica shard就会自动分配过去，然后cluster status就会变成green状态。

1. 索引增删

增：PUT /test\_index?pretty 删：DELETE /test\_index?pretty

然后通过GET /\_cat/health?v来检查

1. CRUD

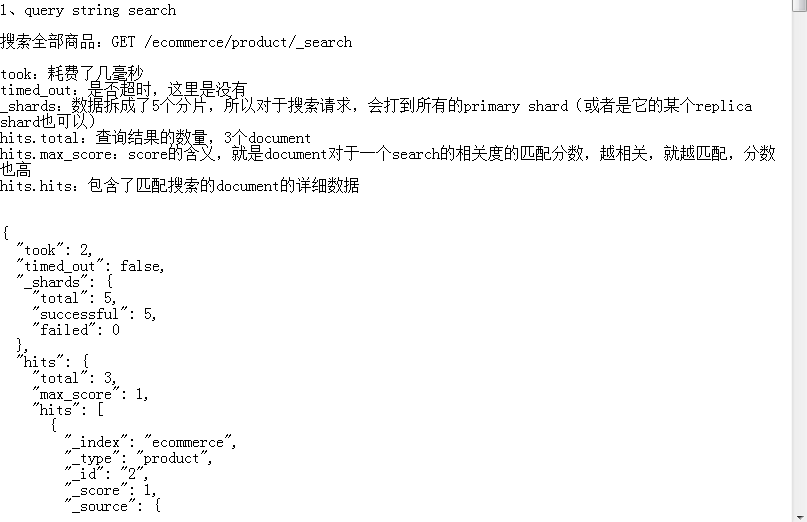
es会自动建立index和type，不需要提前创建，而且es默认会对document每个field都建立倒排索引，让其可以被搜索

* 1. 新增：PUT /index名/type名/id{...}
  2. 查询：GET /index名/type名/id{...}
  3. 更新：
     1. 替换PUT /index名/type名/id{...}

替换方式有一个不好，即使必须带上所有的field，才能去进行信息的修改，否则会把原有的数据给清掉。

* + 1. 修改POST /index名/type名/id/\_update{“doc”:{...}}
  1. 删除：DELETE /index名/type名/id

1. 多种搜索方式
   1. query string search



搜索商品名称中包含yagao的商品，而且按照售价降序排序：

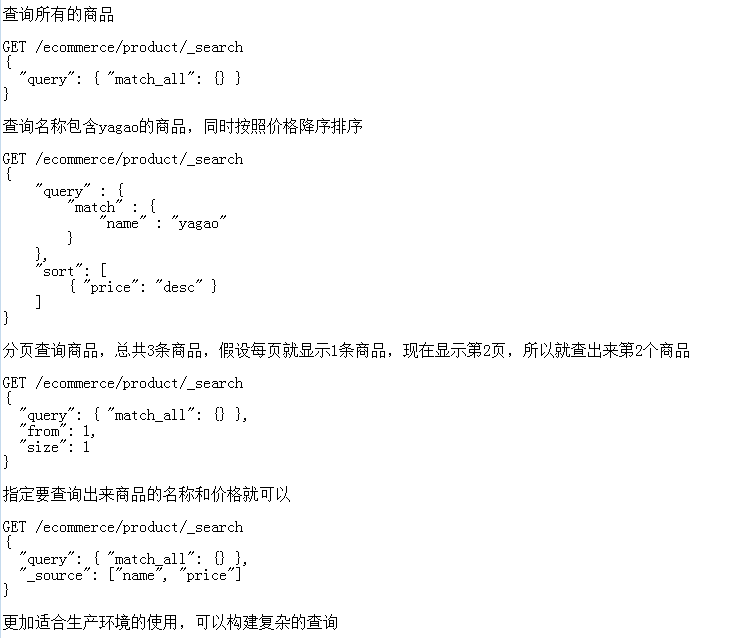
GET /ecommerce/product/\_search?q=name:yagao&sort=price:desc

适用于临时的在命令行使用一些工具，比如curl，快速的发出请求，来检索想要的信息；但是如果查询请求很复杂，是很难去构建的.在生产环境中，几乎很少使用query string search

* 1. query DSL

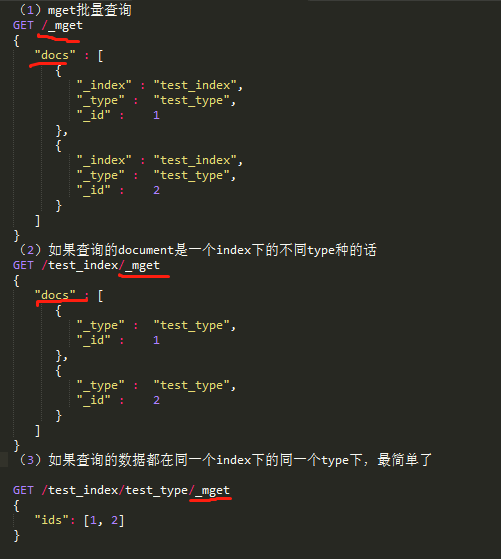
DSL：Domain Specified Language，特定领域的语言

http request body：请求体，可以用json的格式来构建查询语法，比较方便，可以构建各种复杂的语法，比query string search肯定强大多了

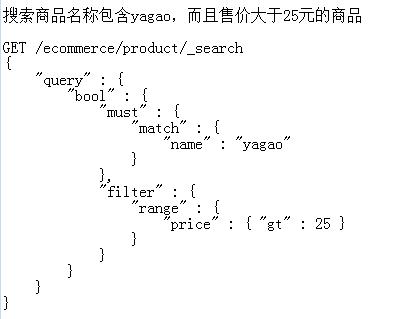


<图解partial update实现原理以及其优点.png>

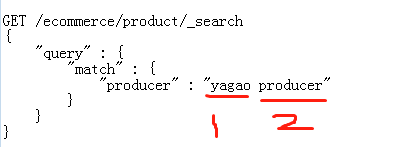
* 1. 批量查询



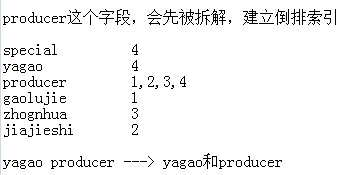
* 1. query filter



* 1. full-text search



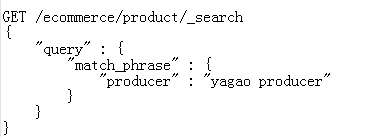
这里有两个词yagao和producer.所以是全文检索。可以看到查出的相关度分数是不一样的。



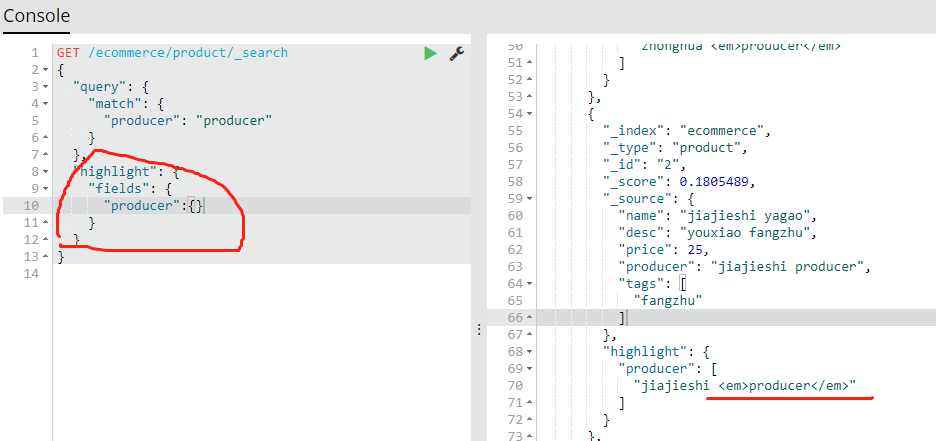
* 1. phrase search（短语搜索）

跟全文检索相对应，相反，全文检索会将输入的搜索串拆解开来，去倒排索引里面去一一匹配，只要能匹配上任意一个拆解后的单词，就可以作为结果返回

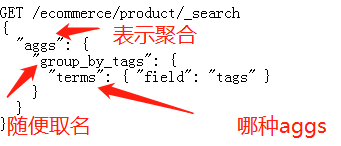
phrase search，要求输入的搜索串，必须在指定的字段文本中，完全包含一模一样的，才可以算匹配，才能作为结果返回



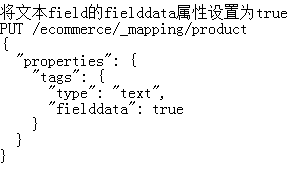
* 1. highlight search（高亮搜索结果）



* 1. 聚合分析



如果提示报错









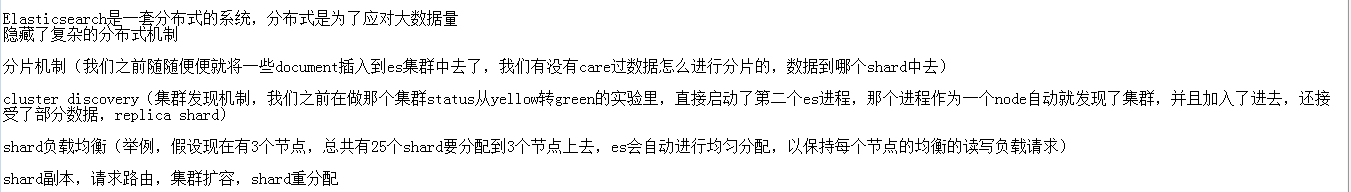


* 1. bulk

1. ES的基础分布式架构

<ES的基础分布式架构.png>

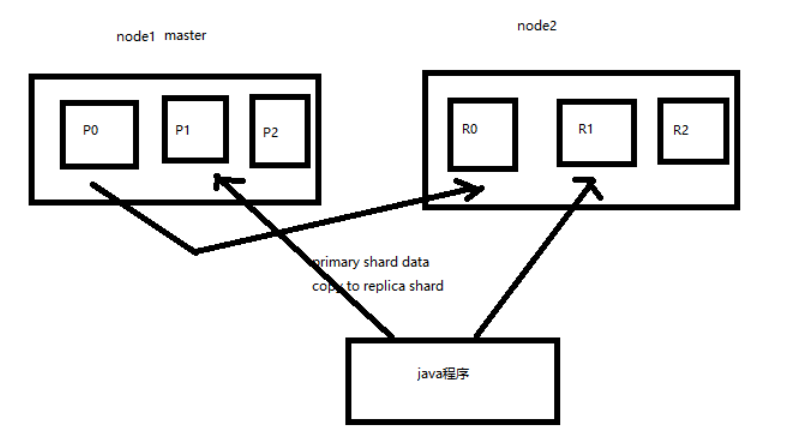
* 1. Elasticsearch对复杂分布式机制的透明隐藏特性



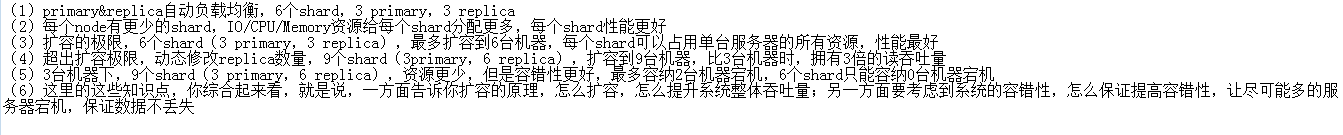
* 1. Elasticsearch的垂直扩容与水平扩容
  2. 增减或减少节点时的数据rebalance
  3. master节点
  4. 节点平等的分布式架构

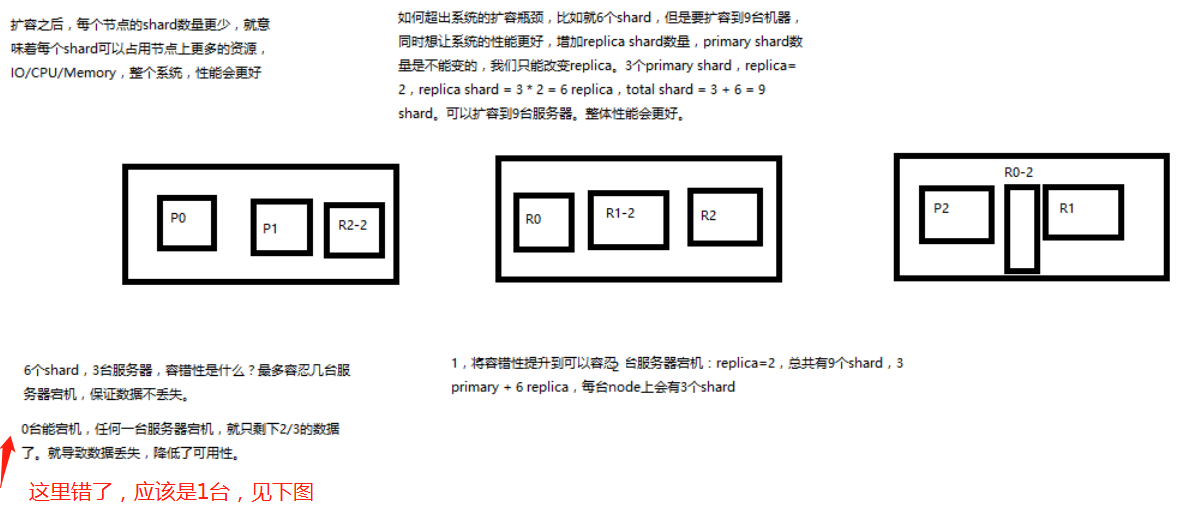
节点对等，每个节点都能接收所有的请求,自动请求路由->响应收集

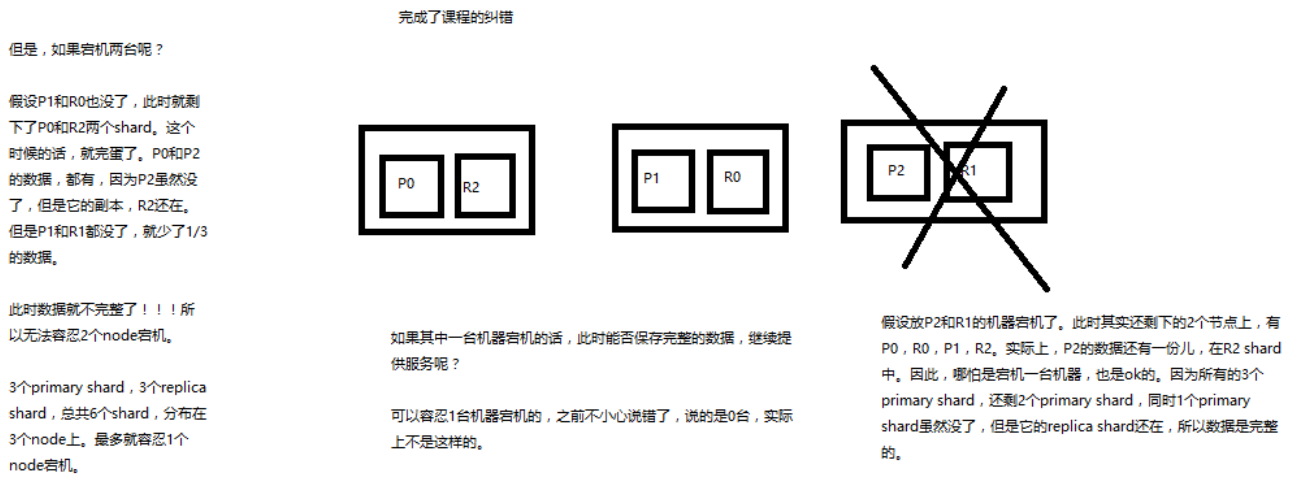
1. shard&replica机制
   1. 单node环境中创建index
   2. 2个node环境下replica shard如何分配



* + 1. replica shard分配：3个primary shard，3个replica shard，1 node
    2. primary ---> replica同步
    3. 读请求：primary/replica
  1. 横向扩容过程，如何超出扩容极限，以及如何提升容错性







* 1. ES容错过程分析

假设9 shard，3 node

master node宕机，自动master选举，red

replica容错：新master将replica提升为primary shard，yellow

重启宕机node，master copy replica到该node，使用原有的shard并同步宕机后的修改，green

<es容错过程分析.png>

1. document核心元数据
   1. 三种元数据
      1. \_index元数据
         1. 代表一个document存放在哪个index中
         2. 类似的数据放在一个索引，非类似的数据放不同索引：product index（包含了所有的商品），sales index（包含了所有的商品销售数据），inventory index（包含了所有库存相关的数据）。如果你把比如product，sales，human resource（employee），全都放在一个大的index里面，比如说company index，不合适的。
         3. index中包含了很多类似的document：类似是什么意思，其实指的就是说，这些document的fields很大一部分是相同的，你说你放了3个document，每个document的fields都完全不一样，这就不是类似了，就不太适合放到一个index里面去了。下图是反例，不应该这样创建index
         4. 索引名称必须是小写的，不能用下划线开头，不能包含逗号

<index如何创建的反例分析.png>

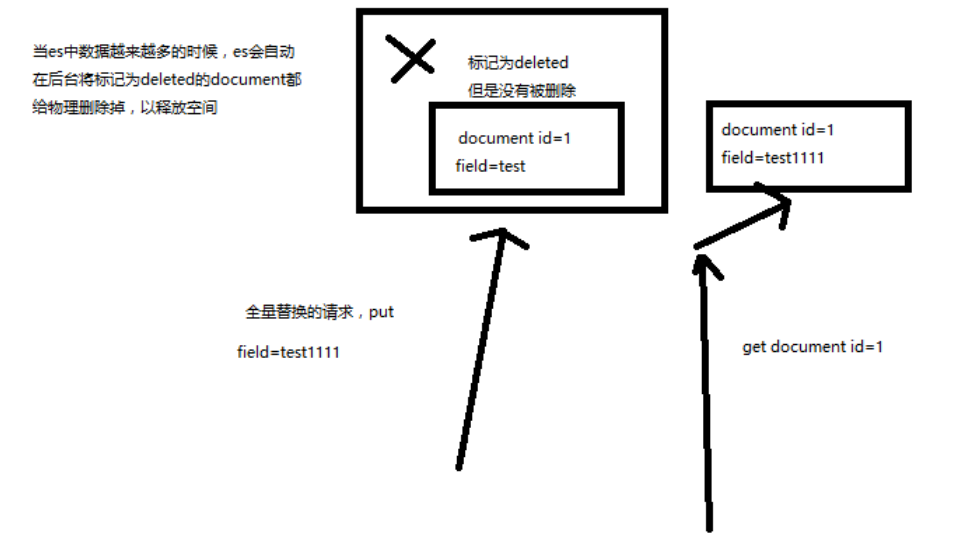
* + 1. \_type元数据
       1. 代表document属于index中的哪个类别（type）
       2. 一个索引通常会划分为多个type，逻辑上对index中有些许不同的几类数据进行分类：因为一批相同的数据，可能有很多相同的fields，但是还是可能会有一些轻微的不同，可能会有少数fields是不一样的，举个例子，就比如说，商品，可能划分为电子商品，生鲜商品，日化商品，等等。
       3. type名称可以是大写或者小写，但是同时不能用下划线开头，不能包含逗号
    2. \_id元数据

代表document的唯一标识，与index和type一起，可以唯一标识和定位一个document.我们可以手动指定document的id（put /index/type/id），也可以不指定，由es自动为我们创建一个id

* + 1. \_source元数据

使用的那个放在request body中的json串，默认情况下，在get的时候，会原封不动的给我们返回回来。如果不想返回那么多，可以在后面加?\_source=属性名或者使用上面查询的querydsl方式

* 1. document id的手动指定与自动生成两种方式
  2. document的全量替换、强制创建以及图解lazy delete机制



* + 1. document的全量替换

document是不可变的，如果要修改document的内容，第一种方式就是全量替换，直接对document重新建立索引，替换里面所有的内容。es会将老的document标记为deleted，然后新增我们给定的一个document，当我们创建越来越多的document的时候，es会在适当的时机在后台自动删除标记为deleted的document

* + 1. document的强制创建

PUT /index/type/id?op\_type=create或者PUT /index/type/id/\_create，一般用后者

* + 1. document的删除

DELETE /index/type/id，不会理解物理删除，只会将其标记为deleted，当数据越来越多的时候，在后台自动删除

1. 并发
   1. 并发冲突

<深度图解剖析Elasticsearch并发冲突问题.png>

* 1. 悲观锁与乐观锁两种并发控制方案

<深度图解剖析悲观锁与乐观锁两种并发控制方案.png>

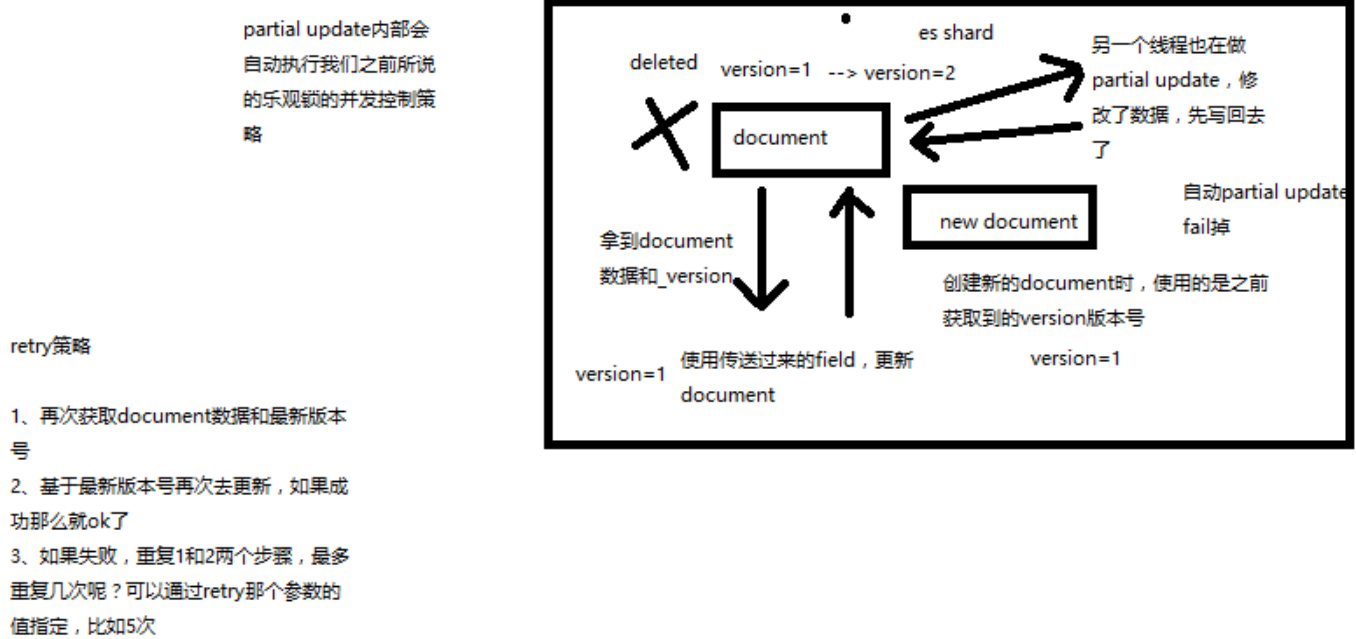
* 1. 基于\_version进行乐观锁并发控制

第一次创建一个document的时候，它的\_version内部版本号就是1；以后，每次对这个document执行修改或者删除操作，都会对这个\_version版本号自动加1；哪怕是删除，也会对这条数据的版本号加1

<图解Elasticsearch内部如何基于_version进行乐观锁并发控制.png>

partial update乐观锁并发控制原理以及相关操作

可以通过post /index/type/id/\_update?retry\_on\_conflict=5&version=6来定义重试次数



* + 1. 基于\_version进行乐观锁并发控制
       1. 先构造一条数据出来

PUT /test\_index/test\_type/7{ "test\_field": "test test"}

* + - 1. 模拟两个客户端，都获取到了同一条数据

GET test\_index/test\_type/7

* + - 1. 其中一个客户端，先更新了一下这个数据，同时带上数据的版本号，确保说，es中的数据的版本号，跟客户端中的数据的版本号是相同的，才能修改

PUT /test\_index/test\_type/7?version=1 { "test\_field": "test client 1"}

* + - 1. 另外一个客户端，尝试基于version=1的数据去进行修改，同样带上version版本号，进行乐观锁的并发控制

PUT /test\_index/test\_type/7?version=1 {"test\_field": "test client 2"}

* + - 1. 在乐观锁成功阻止并发问题之后，尝试正确的完成更新，基于最新的数据和版本号，去进行修改，修改后，带上最新的版本号，可能这个步骤会需要反复执行好几次，才能成功，特别是在多线程并发更新同一条数据很频繁的情况下

GET /test\_index/test\_type/7

PUT /test\_index/test\_type/7?version=2 { "test\_field": "test client 2"}

* + 1. 基于external version进行乐观锁并发控制

?version=1&version\_type=external唯一的区别在于，\_version，只有当你提供的version与es中的\_version一模一样的时候，才可以进行修改，只要不一样，就报错；当version\_type=external的时候，只有当你提供的version比es中的\_version大的时候，才能完成修改

1. document数据路由原理

<document路由原理.png>

* 1. 算法

路由算法：shard = hash(routing) % number\_of\_primary\_shards

举个例子，一个index有3个primary shard，P0，P1，P2

每次增删改查一个document的时候，都会带过来一个routing number，默认就是这个document的\_id（可能是手动指定，也可能是自动生成）routing = \_id

假设\_id=1,会将这个routing值，传入一个hash函数中，产出一个routing值的hash值，hash(routing) = 21,然后将hash函数产出的值对这个index的primary shard的数量求余数，21 % 3 = 0,就决定了，这个document就放在P0上。决定一个document在哪个shard上，最重要的一个值就是routing值，默认是\_id，也可以手动指定，相同的routing值，每次过来，从hash函数中，产出的hash值一定是相同的.

无论hash值是几，无论是什么数字，对number\_of\_primary\_shards求余数，结果一定是在0~number\_of\_primary\_shards-1之间这个范围内的。0,1,2。

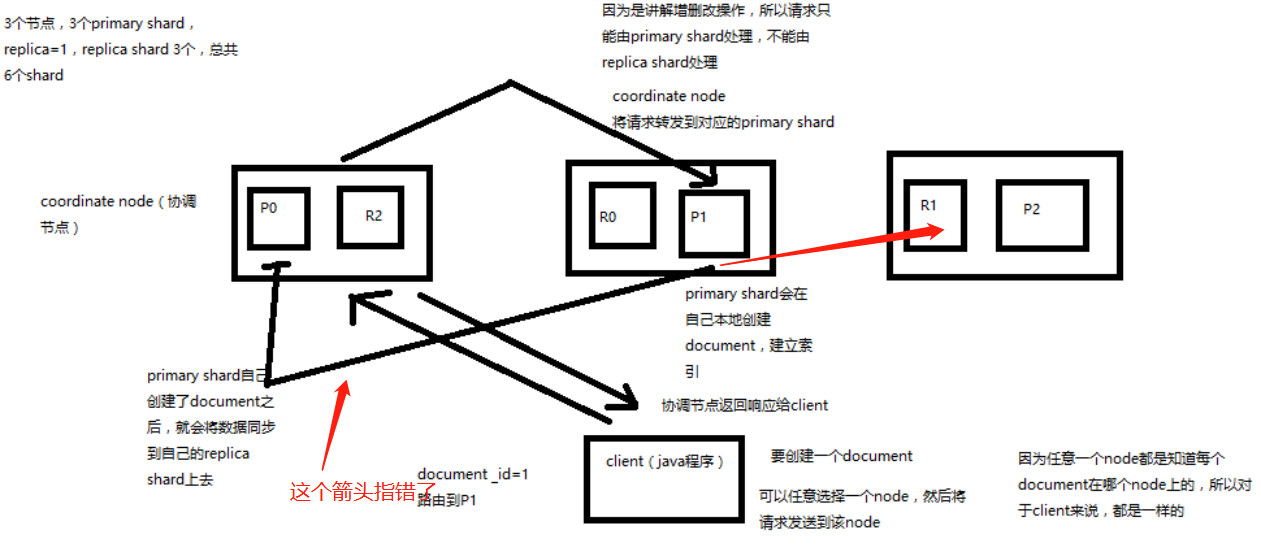
所以说primary shard一旦index建立，是不允许修改的。（见图）。但是replica shard是可以随时修改。

* 1. \_id or custom routing value

默认的routing就是\_id

也可以在发送请求的时候，手动指定一个routing value，比如说put /index/type/id?routing=user\_id。手动指定routing value是很有用的，可以保证说，某一类document一定被路由到一个shard上去，那么在后续进行应用级别的负载均衡，以及提升批量读取的性能的时候，是很有帮助的。

1. document增删改内部原理



（1）客户端选择一个node发送请求过去，这个node就是coordinating node（协调节点）

（2）coordinating node，对document进行路由，将请求转发给对应的node（有primary shard）

（3）实际的node上的primary shard处理请求，然后将数据同步到replica node

（4）coordinating node，如果发现primary node和所有replica node都搞定之后，就返回响应结果给客户端

1. document读请求内部原理

<读请求内部原理.png>

1、客户端发送请求到任意一个node，成为coordinate node

2、coordinate node对document进行路由，将请求转发到对应的node，此时会使用round-robin随机轮询算法，在primary shard以及其所有replica中随机选择一个，让读请求负载均衡

3、接收请求的node返回document给coordinate node

4、coordinate node返回document给客户端

5、特殊情况：document如果还在建立索引过程中，可能只有primary shard有，任何一个replica shard都没有，此时可能会导致无法读取到document，但是document完成索引建立之后，primary shard和replica shard就都有了

1. 写一致性原理以及quorum机制
   1. consistency，one（primary shard），all（all shard），quorum（default）

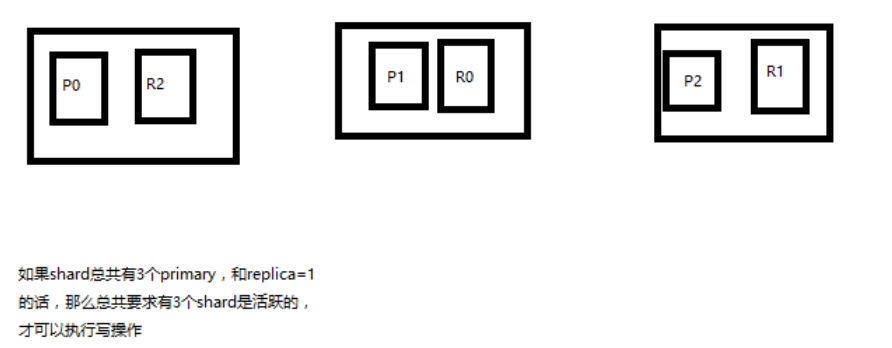
我们在发送任何一个增删改操作的时候，比如说put /index/type/id，都可以带上一个consistency参数，指明我们想要的写一致性是什么？

put /index/type/id?consistency=quorum

one：要求我们这个写操作，只要有一个primary shard是active活跃可用的，就可以执行

all：要求我们这个写操作，必须所有的primary shard和replica shard都是活跃的，才可以执行这个写操作

quorum：默认的值，要求所有的shard中，必须是大部分的shard都是活跃的，可用的，才可以执行这个写操作（见下图）



* 1. quorum机制

写之前必须确保大多数shard都可用，int( (primary + number\_of\_replicas) / 2 ) + 1，当number\_of\_replicas>1时才生效

quroum = int( (primary + number\_of\_replicas) / 2 ) + 1

举个例子，3个primary shard，number\_of\_replicas=1，总共有3 + 3 \* 1 = 6个shard

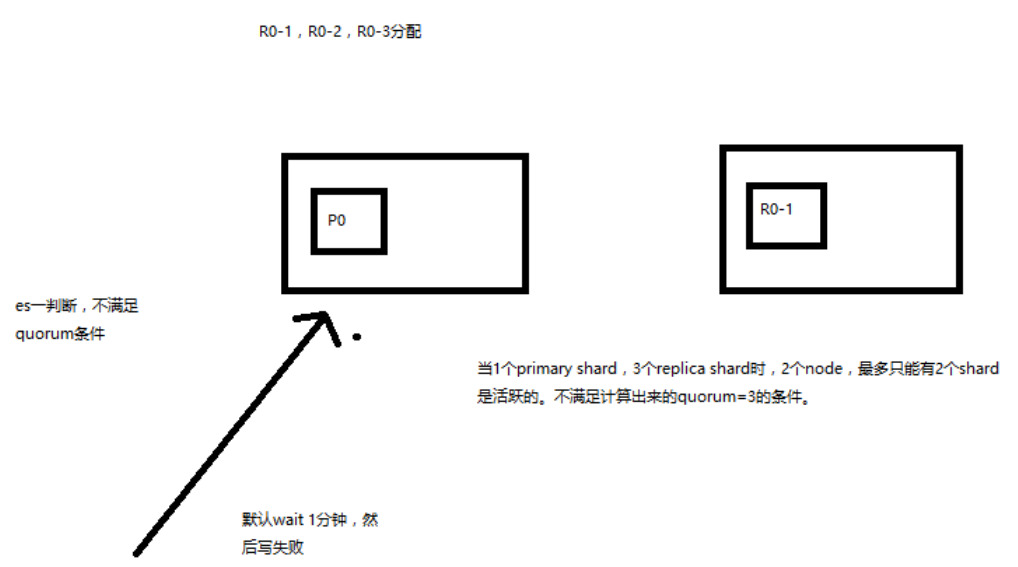
quorum = int( (3 + 1) / 2 ) + 1 = 3

所以，要求6个shard中至少有3个shard是active状态的，才可以执行这个写操作

如果节点数少于quorum数量，可能导致quorum不齐全，进而导致无法执行任何写操作

3个primary shard，replica=1，要求至少3个shard是active，3个shard按照之前学习的shard&replica机制，必须在不同的节点上，如果说只有2台机器的话，是不是有可能出现说，3个shard都没法分配齐全，此时就可能会出现写操作无法执行的情况

1个primary shard，replica=3，quorum=((1 + 3) / 2) + 1 = 3，要求1个primary shard + 3个replica shard = 4个shard，其中必须有3个shard是要处于active状态的。如果这个时候只有2台机器的话，会出现什么情况呢？见下图



es提供了一种特殊的处理场景，就是说当number\_of\_replicas>1时才生效，因为假如说，你就一个primary shard，replica=1，此时就2个shard

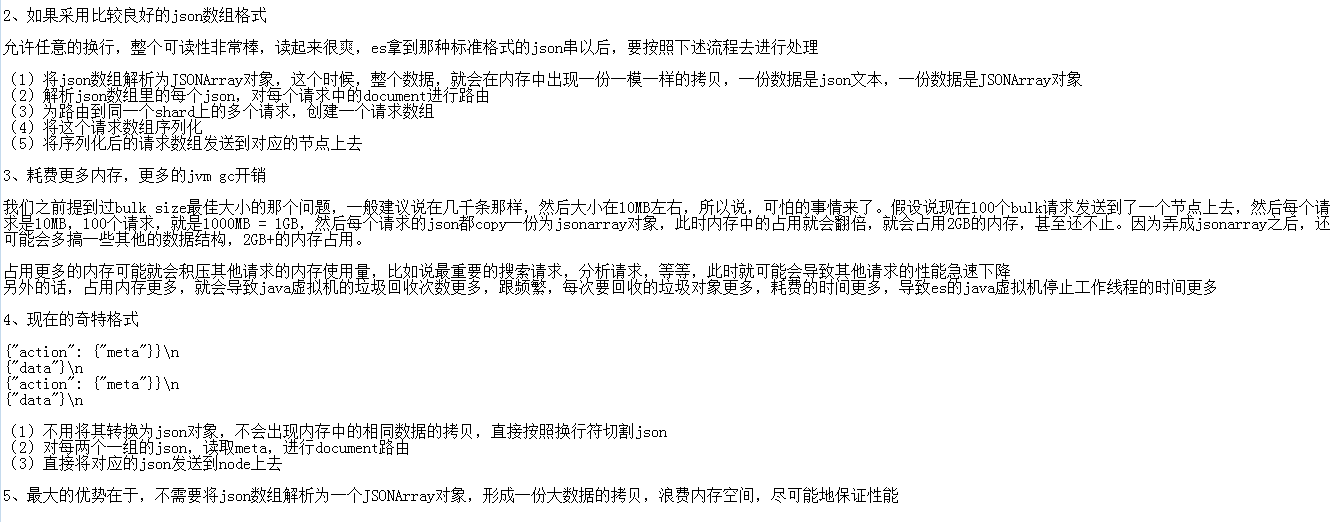
(1 + 1 / 2) + 1 = 2，要求必须有2个shard是活跃的，但是可能就1个node，此时就1个shard是活跃的，如果你不特殊处理的话，导致我们的单节点集群就无法工作

quorum不齐全时，wait，默认1分钟，timeout，100，30s

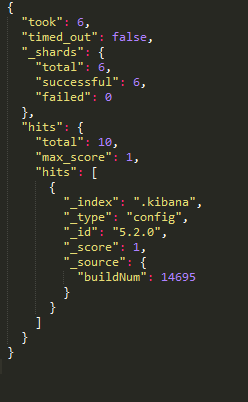
等待期间，期望活跃的shard数量可以增加，最后实在不行，就会timeout

我们其实可以在写操作的时候，加一个timeout参数，比如说put /index/type/id?timeout=30，这个就是说自己去设定quorum不齐全的时候，es的timeout时长，可以缩短，也可以增长

1. bulk api的奇特json格式与底层性能优化关系



1. 搜索



took：整个搜索请求花费了多少毫秒

hits.total：本次搜索，返回了几条结果

hits.max\_score：本次搜索的所有结果中，最大的相关度分数是多少，每一条document对于search的相关度，越相关，\_score分数越大，排位越靠前

hits.hits：默认查询前10条数据，完整数据，\_score降序排序

shards：shards fail的条件（primary和replica全部挂掉），不影响其他shard。默认情况下来说，一个搜索请求，会打到一个index的所有primary shard上去，当然了，每个primary shard都可能会有一个或多个replic shard，所以请求也可以到primary shard的其中一个replica shard上去。

timeout：默认无timeout，latency平衡completeness，手动指定timeout，timeout查询执行机制

timeout=10ms，timeout=1s，timeout=1m

GET /\_search?timeout=10m

<timeout机制详解.png>