bobby · 更新于 2022-11-16

上一节 8. query dsl查询 10. Elasticsearc… 下一节

官方文档

什么是 Mapping?

在一篇文章带你搞定 ElasticSearch 术语中，我们讲到了 Mapping 类似于数据库中的表结构定义 schema，它有以下几个作用：
定义索引中的字段的名称定义字段的数据类型，比如字符串、数字、布尔**字段，倒排索引的相关配置**，比如设置某个字段为不被索引、记录 position 等在 ES 早期版本，一个索引下是可以有多个 Type ，从 7.0 开始，一个索引只有一个 Type，也可以说一个 Type 有一个 Mapping 定义。
在了解了什么是 Mapping 之后，接下来对 Mapping 的设置做下介绍：

Mapping 设置

在创建一个索引的时候，可以对 dynamic 进行设置，可以设成 false、true 或者 strict。

比如一个新的文档，这个文档包含一个字段，当 Dynamic 设置为 true 时，这个文档可以被索引进 ES，这个字段也可以被索引，也就是这个字段可以被搜索，Mapping 也同时被更新；当 dynamic 被设置为 false 时候，存在新增字段的数据写入，该数据可以被索引，但是新增字段被丢弃；当设置成 strict 模式时候，数据写入直接出错。
另外还有 index 参数，用来控制当前字段是否被索引，默认为 true，如果设为 false，则该字段不可被搜索。
参数 index_options 用于控制倒排索引记录的内容，有如下 4 种配置：
doc：只记录 doc id
freqs：记录 doc id 和 term frequencies
positions：记录 doc id、term frequencies 和 term position
offsets：记录 doc id、term frequencies、term position 和 character offects

另外，text 类型默认配置为 positions，其他类型默认为 doc，记录内容越多，占用存储空间越大。

null_value 主要是当字段遇到 null 值时的处理策略，默认为 NULL，即空值，此时 ES 会默认忽略该值，可以通过设定该值设定字段的默认值，另外只有 KeyWord 类型支持设定 null_value。

copy_to 作用是将该字段的值复制到目标字段，实现类似 _all 的作用，它不会出现在 _source 中，只用来搜索。
除了上述介绍的参数，还有许多参数，大家感兴趣的可以在官方文档中进行查看。
在学习了 Mapping 的设置之后，让我们来看下字段的数据类型有哪些吧！

字段数据类型

ES 字段类型类似于 MySQL 中的字段类型，ES 字段类型主要有：核心类型、复杂类型、地理类型以及特殊类型，具体的数据类型如下图所示：

核心类型

从图中可以看出核心类型可以划分为字符串类型、数字类型、日期类型、布尔类型、基于 BASE64 的二进制类型、范围类型。

字符串类型

其中，在 ES 7.x 有两种字符串类型：text 和 keyword，在 ES 5.x 之后 string 类型已经不再支持了。

text 类型适用于需要被全文检索的字段，例如新闻正文、邮件内容等比较长的文字，text 类型会被 Lucene 分词器（Analyzer）处理为一个个词项，并使用 Lucene 倒排索引存储，**text 字段不能被用于排序**，如果需要使用该类型的字段只需要在定义映射时指定 JSON 中对应字段的 type 为 text。
keyword 适合简短、结构化字符串，例如主机名、姓名、商品名称等，**可以用于过滤、排序、聚合检索，也可以用于精确查询。**

数字类型

数字类型分为 long、integer、short、byte、double、float、half_float、scaled_float。
数字类型的字段在满足需求的前提下应当尽量选择范围较小的数据类型，字段长度越短，搜索效率越高，对于浮点数，可以优先考虑使用 scaled_float 类型，该类型可以通过缩放因子来精确浮点数，例如 12.34 可以转换为 1234 来存储。

日期类型

在 ES 中日期可以为以下形式：

格式化的日期字符串，例如 2020-03-17 00:00、2020/03/17
时间戳（和 1970-01-01 00:00:00 UTC 的差值），单位毫秒或者秒即使是格式化的日期字符串，ES 底层依然采用的是时间戳的形式存储。

布尔类型

JSON 文档中同样存在布尔类型，不过 JSON 字符串类型也可以被 ES 转换为布尔类型存储，前提是字符串的取值为 true 或者 false，布尔类型常用于检索中的过滤条件。

二进制类型

二进制类型 binary 接受 BASE64 编码的字符串，默认 store 属性为 false，并且不可以被搜索。

范围类型

范围类型可以用来表达一个数据的区间，可以分为5种：
integer_range、float_range、long_range、double_range 以及 date_range。

复杂类型

复合类型主要有对象类型（object）和嵌套类型（nested）：

对象类型

JSON 字符串允许嵌套对象，一个文档可以嵌套多个、多层对象。可以通过对象类型来存储二级文档，不

实际上 ES 会将其转换为以下格式，并通过 Lucene 存储，即使 name 是 object 类型：

嵌套类型

嵌套类型可以看成是一个特殊的对象类型，可以让对象数组独立检索，例如文档：

username 字段是一个 JSON 数组，并且每个数组对象都是一个 JSON 对象。如果将 username 设置为对象类型，那么 ES 会将其转换为：

可以看出转换后的 JSON 文档中 first 和 last 的关联丢失了，如果尝试搜索 first 为 wu，last 为 xy 的文档，那么成功会检索出上述文档，但是 wu 和 xy 在原 JSON 文档中并不属于同一个 JSON 对象，应当是不匹配的，即检索不出任何结果。
嵌套类型就是为了解决这种问题的，嵌套类型将数组中的每个 JSON 对象作为独立的隐藏文档来存储，每个嵌套的对象都能够独立地被搜索，所以上述案例中虽然表面上只有 1 个文档，但实际上是存储了 4 个文档。

地理类型

地理类型字段分为两种：经纬度类型和地理区域类型：

经纬度类型

经纬度类型字段（geo_point）可以存储经纬度相关信息，通过地理类型的字段，可以用来实现诸如查找在指定地理区域内相关的文档、根据距离排序、根据地理位置修改评分规则等需求。

地理区域类型

经纬度类型可以表达一个点，而 geo_shape 类型可以表达一块地理区域，区域的形状可以是任意多边形，也可以是点、线、面、多点、多线、多面等几何类型。

特殊类型

特殊类型包括 IP 类型、过滤器类型、Join 类型、别名类型等，在这里简单介绍下 IP 类型和 Join 类型，其他特殊类型可以查看官方文档。

IP 类型

IP 类型的字段可以用来存储 IPv4 或者 IPv6 地址，如果需要存储 IP 类型的字段，需要手动定义映射：

Join 类型

Join 类型是 ES 6.x 引入的类型，以取代淘汰的 _parent 元字段，用来实现文档的一对一、一对多的关系，主要用来做父子查询。
Join 类型的 Mapping 如下：

其中，my_join_field 为 Join 类型字段的名称；relations 指定关系：question 是 answer 的父类。
例如定义一个 ID 为 1 的父文档：

接下来定义一个子文档，该文档指定了父文档 ID 为 1：

再了解完字段数据类型后，再让我们看下什么是 Dynamic Mapping?

什么是 Dynamic Mapping?

Dynamic Mapping 机制使我们不需要手动定义 Mapping，ES 会**自动根据文档信息来判断字段合适的类型**，但是有时候也会推算的不对，比如地理位置信息有可能会判断为 Text，当类型如果设置不对时，会导致一些功能无法正常工作，比如 Range 查询。

类型自动识别

ES 类型的自动识别是基于 JSON 的格式，如果输入的是 JSON 是字符串且格式为日期格式，ES 会自动设置成 Date 类型；当输入的字符串是数字的时候，ES 默认会当成字符串来处理，可以通过设置来转换成合适的类型；如果输入的是 Text 字段的时候，ES 会自动增加 keyword 子字段，还有一些自动识别如下图所示：

下面我们通过一个例子是看看是怎么类型自动识别的，输入如下请求，创建索引：

然后使用 GET /mapping_test/_mapping 查看，结果如下图所示：

可以从结果中看出，ES 会根据文档信息自动推算出合适的类型。
哦豁，万一我想修改 Mapping 的字段类型，能否更改呢？让我们分以下两种情况来探究下：

修改 Mapping 字段类型?

如果是新增加的字段，根据 Dynamic 的设置分为以下三种状况：

当 Dynamic 设置为 true 时，一旦有新增字段的文档写入，Mapping 也同时被更新。
当 Dynamic 设置为 false 时，索引的 Mapping 是不会被更新的，新增字段的数据无法被索引，也就是无法被搜索，但是信息会出现在 _source 中。
当 Dynamic 设置为 strict 时，文档写入会失败。

另外一种是字段已经存在，这种情况下，ES 是不允许修改字段的类型的，因为 ES 是根据 Lucene 实现的倒排索引，一旦生成后就不允许修改，如果希望改变字段类型，必须使用 Reindex API 重建索引。
不能修改的原因是如果修改了字段的数据类型，会导致已被索引的无法被搜索，但是如果是增加新的字段，就不会有这样的影响。
**

8. query dsl查询 ◂ 上一节　　下一节 ▸ 10. Elasticsearch Analyze

我要提出意见反馈