



BIG DATA

# Elasticsearch

## 介紹與經驗分享

大數據股份有限公司

VPE Johnny 蔡協哲

2024.06.19

# 網路溫度計

利用大數據分析的網站，主要透過《KEYPO大數據關鍵引擎》進行輿情分析，發表每日調查報告和口碑聲量排名。

該網站提供企業品牌、政治、科技、娛樂等多領域的關鍵數據和產業洞察，幫助用戶了解網路趨勢及輿論動態。

其特點在於中立、具公信力的數據分析，為各界提供精準的網路聲量和趨勢報告



## DailyView 每日調查



巨人交流賽3.8萬人進場破紀錄！大巨蛋啟用後前十大熱門比賽揭曉

臺北大巨蛋2023年12月正式啟用後，已辦過不少場討論度很高的比賽，究竟是哪幾場比賽最受到關注，帶你一文了解。

## Focus 特別企劃



《網路溫度計》10歲生日快樂 眾星雲集來祝賀

2024亞太永續博覽會起跑 簡又新大讚Z世代  
Z世代線上購物行為調查

## KEYPO TOP3 口碑聲量排行TOP3

宅經濟 / 生鮮網購

看完整排名



1 台灣好農

# 輿情數據分析系統

集結最新社群頻道來源及多國語系，每月處理超過1,000億中文數據，40億筆資料查詢最快2秒。

全新介面升級結合最新生成式 AI 技術，率全台輿情分析系統之先，首創「ChatGPT 智能分析功能」，一鍵產出洞察分析速報，提供有效洞見。

輿情即時通知功能，自行設定關注焦點及主題，24 小時掌握社群情報，最快15分鐘即時資訊到手。



# 大綱

01 Elasticsearch 介紹

02 基本知識與欄位型態介紹

03 資料寫入

04 資料查詢

05 實作與Demo

# Elasticsearch

開源的分散式即時全文搜尋和分析引擎，且可使用無結構化的 JSON 為輸入的文件。

## 全文搜尋

提供一種非常強大的全文搜尋功能，基於 Lucene 提供的搜尋功能進行了封裝

## 分散搜尋

將數據分散存儲在多個節點的索引(index)中，這允許在多個節點上同時進行搜尋，從而在大量數據中實現快速搜尋

## 即時分析

一個大規模數據的分析系統，能夠在短時間內對大量數據進行分析並得出結果

## 穩定性 擴展性

穩定地存儲大量數據，並且能夠從一個節點擴展到數百個節點，同時保持高效的操作性



# 叢集名詞說明

- **Cluster**

多台Server組成叢集

- **Node**

叢集中的每一台 Server

- **Index**

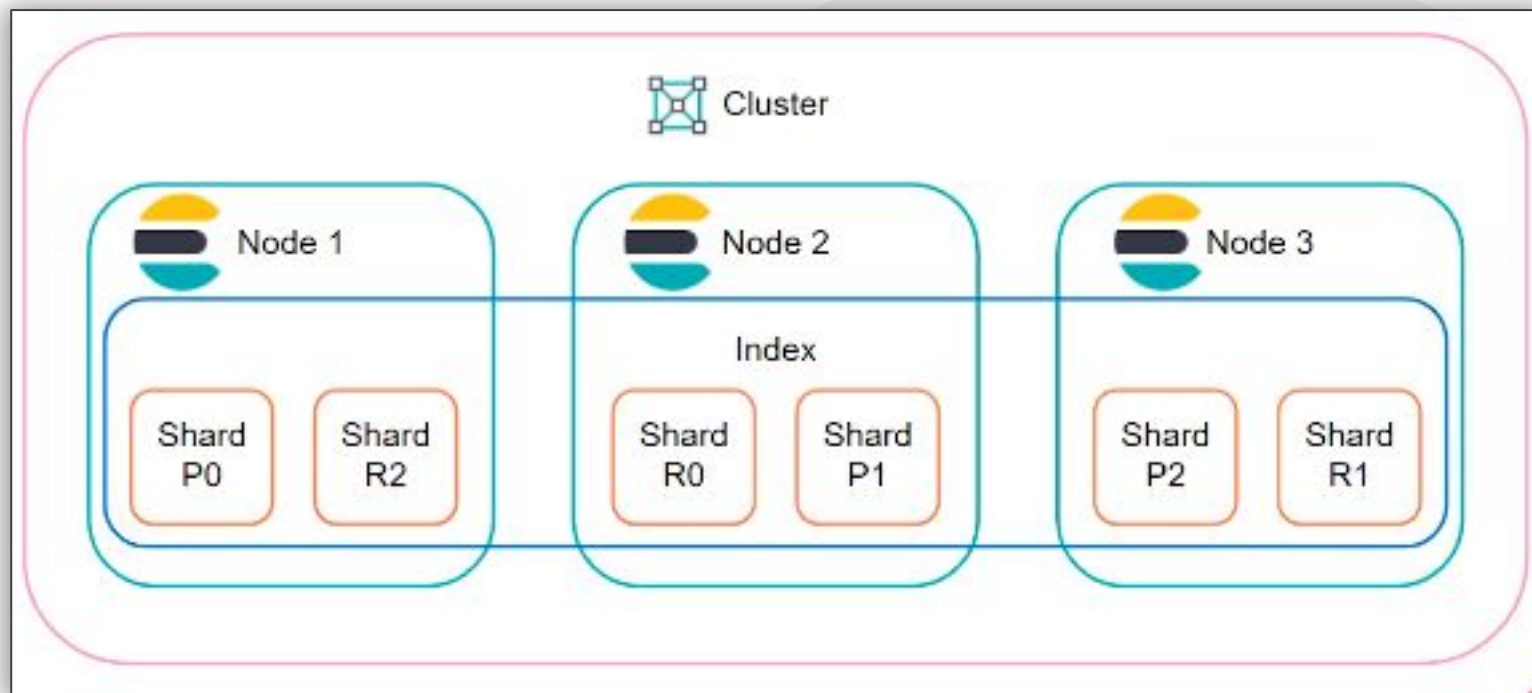
類似RDB的Database

- **Type**

類似RDB的Table

- **Document**

儲存的一筆一筆的資料



# 資料名詞說明

- **Type**

- ES 7以後已移除Type, 因影響儲存效率  
(參考)
- ES 7、8使用 \_doc 來當作過渡期的用法

- **Document**

- 儲存的資料, json格式

```
GET users/_doc/1
```

```
{  
  "_index" : "users",  
  "_type" : "_doc",  
  "_id" : "1",  
  "_version" : 32,  
  "_seq_no" : 36,  
  "_primary_term" : 1,  
  "found" : true,  
  "_source" : {  
    "firstName" : "Jack",  
    "lastName" : "Johnson",  
    "age" : 39,  
    "tags" : [  
      "guitar",  
      "skateboard"  
    ]  
  }  
}
```



```
{
  "_index": "sm-000002",
  "_type": "post",
  "_id": "e4e021ba2fc97eafb260dd1e5d054501",
  "_score": 0.0,
  "_source":
  {
    "doc.id": "null",
    "post.id": "e4e021ba2fc97eafb260dd1e5d054501",
    "post.type": "post",
    "post.poster": "東風衛視",
    "url": "https://www.youtube.com/watch?v=0WNkSBTYvWY",
    "title": "交往10年步入禮堂！美女心理師驚人真面目，婚前財產「全上繳男友」真相震撼馬在勤律師 | 每周精選",
    "posttime": "2023-05-28 19:00",
    "fetchedtime": "2023-05-31 11:22",
    "author.id": "東風衛視",
    "author.name": "東風衛視",
    "content": "交往十年準備結婚財產全交給另一半？買房一人一半但是地點他決定！",
    "count.like": 21,
    "count.comment": 0,
    "count.share": 0,
    "count.view": 4892,
    "source": "YT",
    "channel": "UU2ATMznIeVwWJUH8Fx4KDWa",
    "category": "UU2ATMznIeVwWJUH8Fx4KDWa",
    "fid": "UU2ATMznIeVwWJUH8Fx4KDWa",
    "sentiment": "Neutral",
    "positive_score": 0.9997681840095292,
    "negative_score": 7.93103550313756e-09,
    "site.name": "YOUTUBE",
    "site.type": "SM",
    "site.channel": "東風衛視",
    "ad_score": 0.0,
    "page_tags": "",
    "title_tags": "交往 10 步入 禮堂 美女 心理 驚人真面目 財產 男友 上繳 真相 震撼 律師 精選",
    "content_tags": " 結婚 準備 交往 財產 交給 一半 買房 一人 ",
    "hash_tags": "蘇予昕 結婚 交往",
    "site_channel_category": "",
    "image_url": "https://img.youtube.com/vi/0WNkSBTYvWY/maxresdefault.jpg",
  }
}
```



# 儲存架構

- **Index**
  - 多台Server組成叢集
- **Shard**
  - 每個shard分片是一个Lucene實體
- **Segment**
  - Lucene内部的mini-index
  - 結構：
    - 正排索引(forward index)
    - 倒排索引(Inverted Index)
    - Stored Fields
    - Document Values
    - Cache



# Document如何寫入Elasticsearch?

1. Doc寫入Elasticsearch時, 首先寫入 **記憶體緩衝區**
2. Default每隔1秒, 將**記憶體緩衝區**中的數據寫入一個新的 **Segment文件**中, 並進入Filesystem cache(同時**清空記憶體緩衝區**), 這個過程稱為Refresh
3. 每個Segment包含了**正排索引**與**倒排索引**, 倒排索引用於搜索, 正排索引用於排序、聚合以及回傳文件的原始欄位 值

\* 只有經歷了refresh操作之後, 數據才能變成可檢索的

\* Elasticsearch有一個背景程序專門負責 Segment的合併, 定期執行 merge操作, 將多個小 segment文件合併成一個Segment

# Mapping

定義Index Document各欄位的類型與 內容如何被搜尋

# 基本的 Mapping

- Integer
  - 整數
- Keyword
  - 字串
  - 只支援完全匹配搜尋
- Text
  - 內容斷詞後index儲存,
  - 可進行片段內容搜尋
- 更多Mapping型態可參考  
[參考官方文件](#)

```
PUT /my-index-000001
{
  "mappings": {
    "properties": {
      "age":    { "type": "integer" }, 1
      "email":  { "type": "keyword" }, 2
      "name":   { "type": "text" }     3
    }
  }
}
```

# 多層次的設定

- 如果是多層型態
- 在搜尋時欄位可下如『count.comment』

```
"count": {
  "properties": {
    "angry": {
      "type": "integer"
    },
    "comment": {
      "type": "integer"
    },
    "dislike": {
      "type": "integer"
    },
    "haha": {
      "type": "integer"
    },
    "like": {
      "type": "integer"
    },
    "love": {
      "type": "integer"
    },
    "rating": {
      "type": "float"
    },
    "reactions": {
      "type": "integer"
    },
    "sad": {
      "type": "integer"
    },
    "share": {
      "type": "integer"
    },
    "thankful": {
      "type": "integer"
    },
    "view": {
      "type": "integer"
    },
    "wow": {
      "type": "integer"
    }
  }
}
```



# 時間格式

- Date type中可設定format格式
- 寫入之資料字串就必須符合該格式

```
"created_time": {  
  "type": "date",  
  "format": "yyyy-MM-dd HH:mm"  
},  
"id": 1
```

# 多層不同型態應用

- Name中又包含Keyword與Text型態
- Text使用外掛jieba套件進行斷詞，其斷詞結果供搜尋使用
- 更多Analyzer
  - 參考官方文件

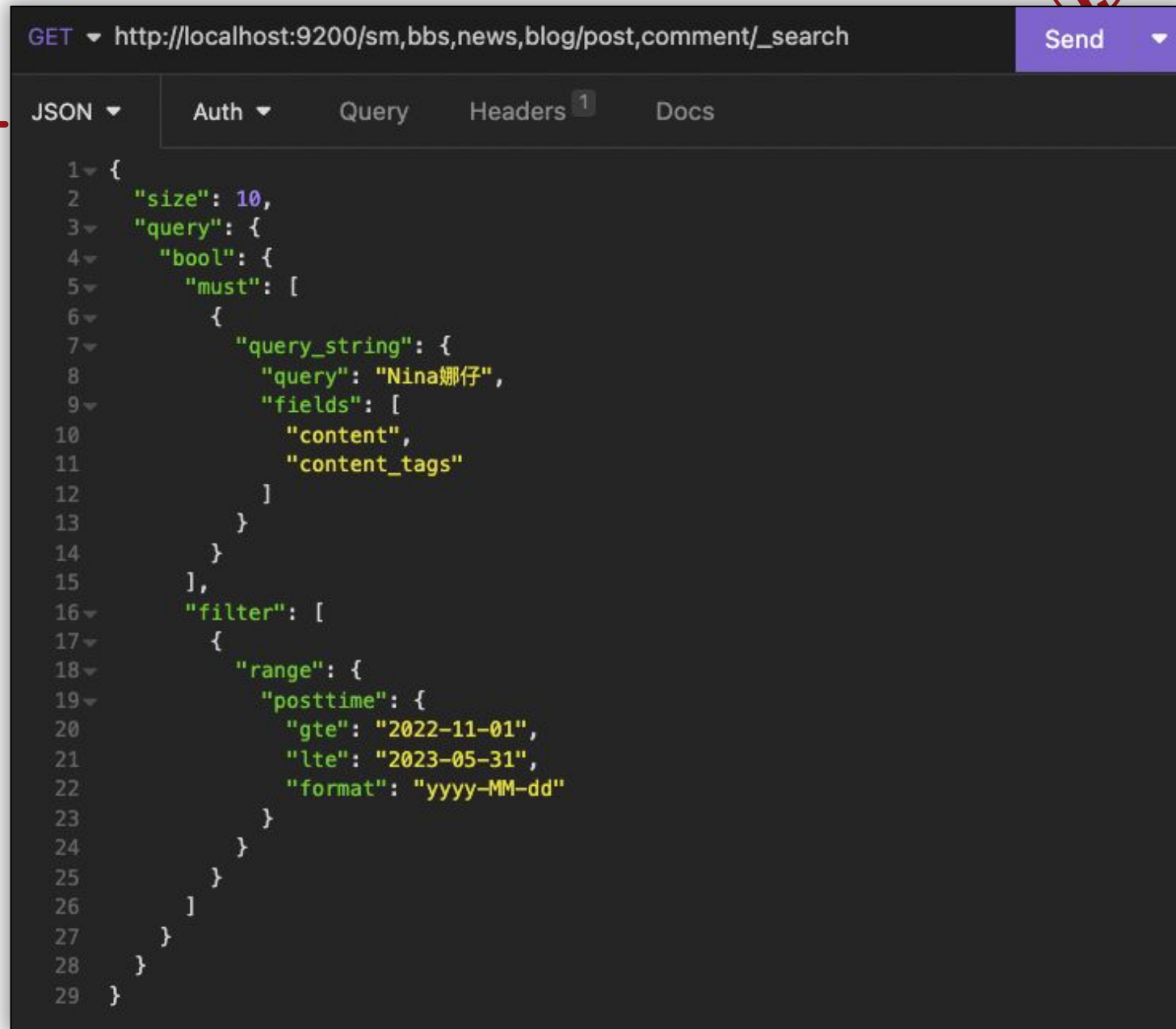
```
"author": { -
  "properties": { -
    "id": { -
      "type": "keyword"
    },
    "name": { -
      "type": "text",
      "norms": false,
      "eager_global_ordinals": true,
      "fields": { -
        "raw": { -
          "type": "keyword"
        }
      }
    },
    "analyzer": "jieba_analyzer",
    "fielddata": true,
    "fielddata_frequency_filter": { -
      "min": 0.001,
      "max": 0.1,
      "min_segment_size": 5000
    }
  }
},
},
},
```

# 常見資料查詢

# 基本查詢

- bool
  - must、must\_not
  - filter
- query\_string
- term、terms
- 更多query參考

官方文件



The screenshot shows a REST client interface with a GET request to `http://localhost:9200/sm,bbs,news,blog/post,comment/_search`. The JSON body is as follows:

```
1 {
2   "size": 10,
3   "query": {
4     "bool": {
5       "must": [
6         {
7           "query_string": {
8             "query": "Nina娜仔",
9             "fields": [
10              "content",
11              "content_tags"
12            ]
13          }
14        }
15      ],
16      "filter": [
17        {
18          "range": {
19            "posttime": {
20              "gte": "2022-11-01",
21              "lte": "2023-05-31",
22              "format": "yyyy-MM-dd"
23            }
24          }
25        }
26      ]
27    }
28  }
29 }
```

# 基本查詢

- bool
  - must、must\_not
  - filter
- query\_string
- term、terms
- 更多query參考

官方文件

```
"filter": [  
  {  
    "terms": {  
      "url": [  
        "https://www.facebook.com/56712367053_750963853061038",  
        "https://www.facebook.com/56712367053_780537713436985",  
        "https://www.facebook.com/56712367053_748811703276253",  
        "https://www.instagram.com/p/CoCJ6TmrfUB",  
        "https://www.instagram.com/p/CqxthQQLpne",  
        "https://www.instagram.com/p/Com6yyhDX7w",  
        "https://www.instagram.com/p/CoearunPGor",  
        "https://www.youtube.com/watch?v=8BLcpTkapXo",  
        "https://www.youtube.com/watch?v=VbP_S6fhcU4"  
      ]  
    }  
  },  
]
```

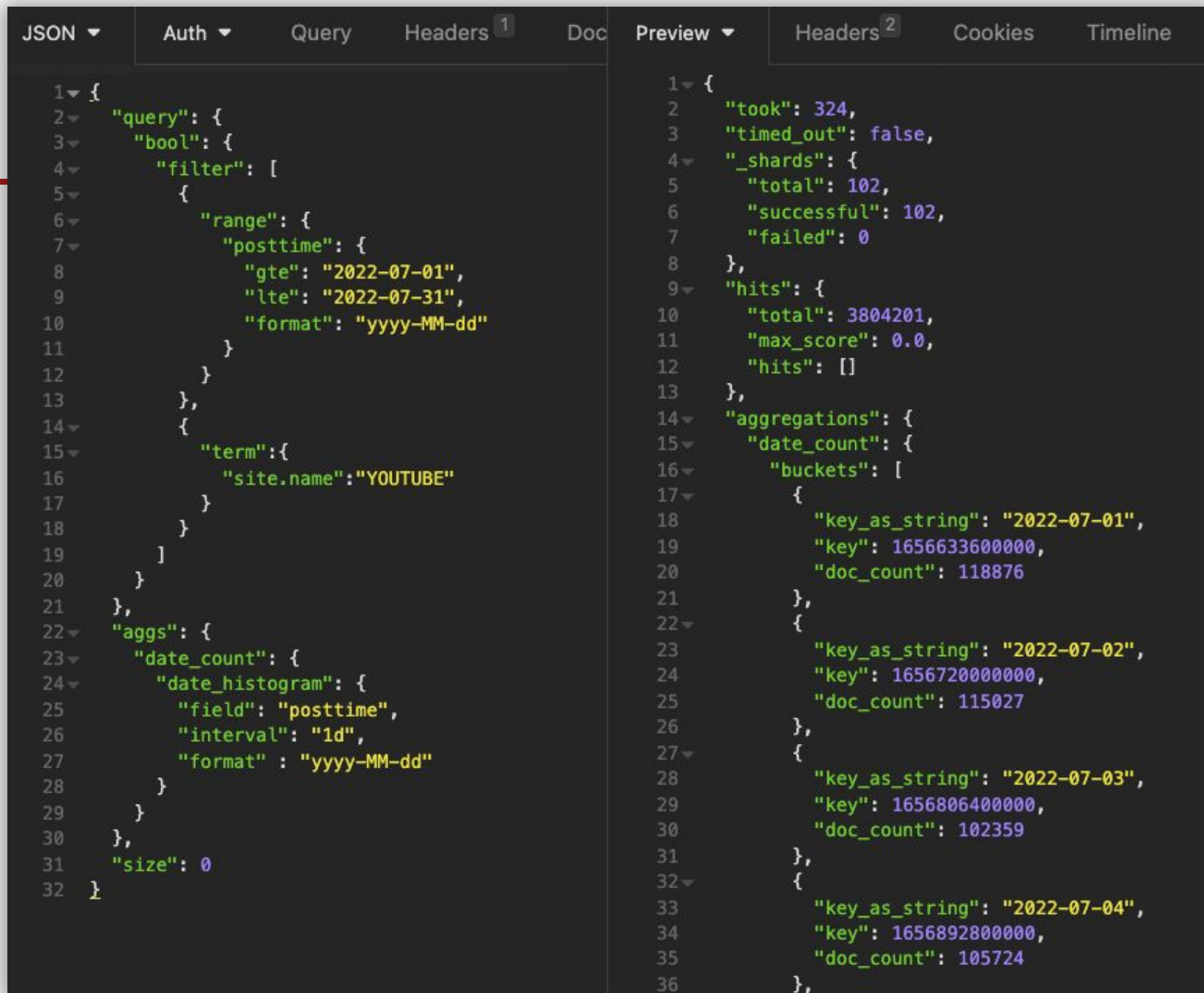
```
"filter": [  
  {  
    "term": {  
      "post.id": "e4e021ba2fc97eafb260dd1e5d054501"  
    }  
  }  
]
```



# 聚合查詢

- aggs
- 日期聚合 date\_histogram
  - field
  - interval
  - format
- 更多 aggs query 參考

官方文件



The screenshot displays a REST client interface with two panels. The left panel shows the JSON query, and the right panel shows the JSON response.

**Query (Left Panel):**

```
1 {
2   "query": {
3     "bool": {
4       "filter": [
5         {
6           "range": {
7             "posttime": {
8               "gte": "2022-07-01",
9               "lte": "2022-07-31",
10              "format": "yyyy-MM-dd"
11            }
12          }
13        },
14        {
15          "term": {
16            "site.name": "YOUTUBE"
17          }
18        }
19      ]
20    }
21  },
22  "aggs": {
23    "date_count": {
24      "date_histogram": {
25        "field": "posttime",
26        "interval": "1d",
27        "format": "yyyy-MM-dd"
28      }
29    }
30  },
31  "size": 0
32 }
```

**Response (Right Panel):**

```
1 {
2   "took": 324,
3   "timed_out": false,
4   "_shards": {
5     "total": 102,
6     "successful": 102,
7     "failed": 0
8   },
9   "hits": {
10    "total": 3804201,
11    "max_score": 0.0,
12    "hits": []
13  },
14  "aggregations": {
15    "date_count": {
16      "buckets": [
17        {
18          "key_as_string": "2022-07-01",
19          "key": 1656633600000,
20          "doc_count": 118876
21        },
22        {
23          "key_as_string": "2022-07-02",
24          "key": 1656720000000,
25          "doc_count": 115027
26        },
27        {
28          "key_as_string": "2022-07-03",
29          "key": 1656806400000,
30          "doc_count": 102359
31        },
32        {
33          "key_as_string": "2022-07-04",
34          "key": 1656892800000,
35          "doc_count": 105724
36        }
37      ]
38    }
39  }
40 }
```

# 聚合查詢

- Keywords型態欄位聚合

- field
- size

- 更多aggs query參考

官方文件

```
GET http://localhost:9200/sm,bbs,new Send 200 OK 2.53 s 1364 B
JSON Auth Query Headers 1 Do Preview 2 Headers Cookies Timeline
{
  "query_string": {
    "query": "\"小\"",
    "fields": [
      "author.name"
    ]
  },
  "filter": [
    {
      "term": {
        "site.name": "Dcard論壇"
      }
    },
    {
      "range": {
        "posttime": {
          "gte": "2022-01-01",
          "lte": "2022-08-01",
          "format": "yyyy-MM-dd"
        }
      }
    }
  ],
  "aggs": {
    "channel": {
      "terms": {
        "field": "author.name.raw",
        "size": 20
      }
    }
  }
}
```

```
{
  "took": 2505,
  "timed_out": false,
  "_shards": {
    "total": 88,
    "successful": 88,
    "failed": 0
  },
  "hits": {
    "total": 14063,
    "max_score": 0.0,
    "hits": []
  },
  "aggregations": {
    "channel": {
      "doc_count_error_upper_bound": 32,
      "sum_other_doc_count": 8504,
      "buckets": [
        {
          "key": "打工小天使 dcardparttime",
          "doc_count": 1853
        },
        {
          "key": "小廢物4我 mewithoutyou",
          "doc_count": 715
        },
        {
          "key": "Demo小天使 dcarddemoangel",
          "doc_count": 510
        },
        {
          "key": "遊戲小天使 gameangel",
          "doc_count": 366
        },
        {
          "key": "實習小天使 dcardintern",

```

# 聚合查詢

GET http://localhost:9200/sm/post/\_search

Send 200 OK 7.9 s 731 B

JSON Auth Query Headers<sup>1</sup> Docs

```
https://www.youtube.com/channel/UCz_rK02a3hnmM00USmFQo2w\",
  \"fields\": [
    \"title\",
    \"content\"
  ]
},
{
  \"term\": {
    \"sentiment\": \"Positive\"
  }
}
]
},
{
  \"aggs\": {
    \"channel\": {
      \"terms\": {
        \"script\": \"doc['site.name'].value+'/' + doc['site.channel.raw'].value\"
      }
    }
  },
  \"size\": 0
}
```

Preview Headers<sup>2</sup> Cookies Timeline

```
13 },
14 \"aggregations\": {
15   \"channel\": {
16     \"doc_count_error_upper_bound\": 194,
17     \"sum_other_doc_count\": 15368,
18     \"buckets\": [
19       {
20         \"key\": \"YOUTUBE/TVBS選新聞\",
21         \"doc_count\": 19409
22       },
23       {
24         \"key\": \"YOUTUBE/中天新聞\",
25         \"doc_count\": 9569
26       },
27       {
28         \"key\": \"YOUTUBE/TVBS NEWS\",
29         \"doc_count\": 8430
30       },
31       {
32         \"key\": \"YOUTUBE/寰宇新聞 頻道\",
33         \"doc_count\": 6371
34       },
35       {
36         \"key\": \"YOUTUBE/中視新聞\",
37         \"doc_count\": 6350
38       },
39       {
40         \"key\": \"YOUTUBE/三立新聞網SETN\",
41         \"doc_count\": 6165
42       },
43       {
44         \"key\": \"YOUTUBE/中時新聞網\",
45         \"doc_count\": 4055
46       },
47       {
48         \"key\": \"YOUTUBE/快點TV\",
```

```
\"aggs\": {
  \"channel\": {
    \"terms\": {
      \"script\": \"doc['site.name'].value+'/' + doc['site.channel.raw'].value\"
    }
  }
},
```

# 實作範例與 Demo

Python套件：

1. elasticsearch
2. elasticsearch-dsl



# 資料寫入

- 寫入分成index、bulk API
- 更新doc有update和upsert

```
# Connect to Elasticsearch
es = Elasticsearch(hosts=["http://localhost:9200"])

# Document data
docs = [
    {
        "_index": "post",
        "_id": "1",
        "_source": {
            "post": {"id": "1", "type": "post", "poster": "小資yp投資理財筆記"},
            "site": {"name": "FACEBOOK", "channel": "小資yp投資理財筆記"},
            "title": "最近的AI話題非常火熱",
            "posttime": "2024-06-15 09:23",
            "author": {"id": "user1", "name": "小資yp投資理財筆記"},
            "content": "最近的AI話題非常火熱，從股票市場中輝達(NVIDIA)的表現..."
        }
    },
    { ...
    { ...
]
```

```
# Use bulk API to add documents
helpers.bulk(es, docs)
```



# 資料查詢

- 可使用elasticsearch套件發送json格式查詢
- 或使用elasticsearch-dsl使用物件導向程式查詢

```
# Connect to Elasticsearch
es = Elasticsearch(hosts=["http://localhost:9200"])

# Query string
query_string = 'NVIDIA OR (AI AND "黃仁勳")'

# Execute the search query
response = es.search(
    index="post", body={"query": {"query_string": {"query": query_string}}}
)

# Print the results
for hit in response["hits"]["hits"]:
    print(
        f"ID: {hit['_id']}, Title: {hit['_source']['title']}, Content: {hit['_source']['content']}"
    )
```



# FAQ