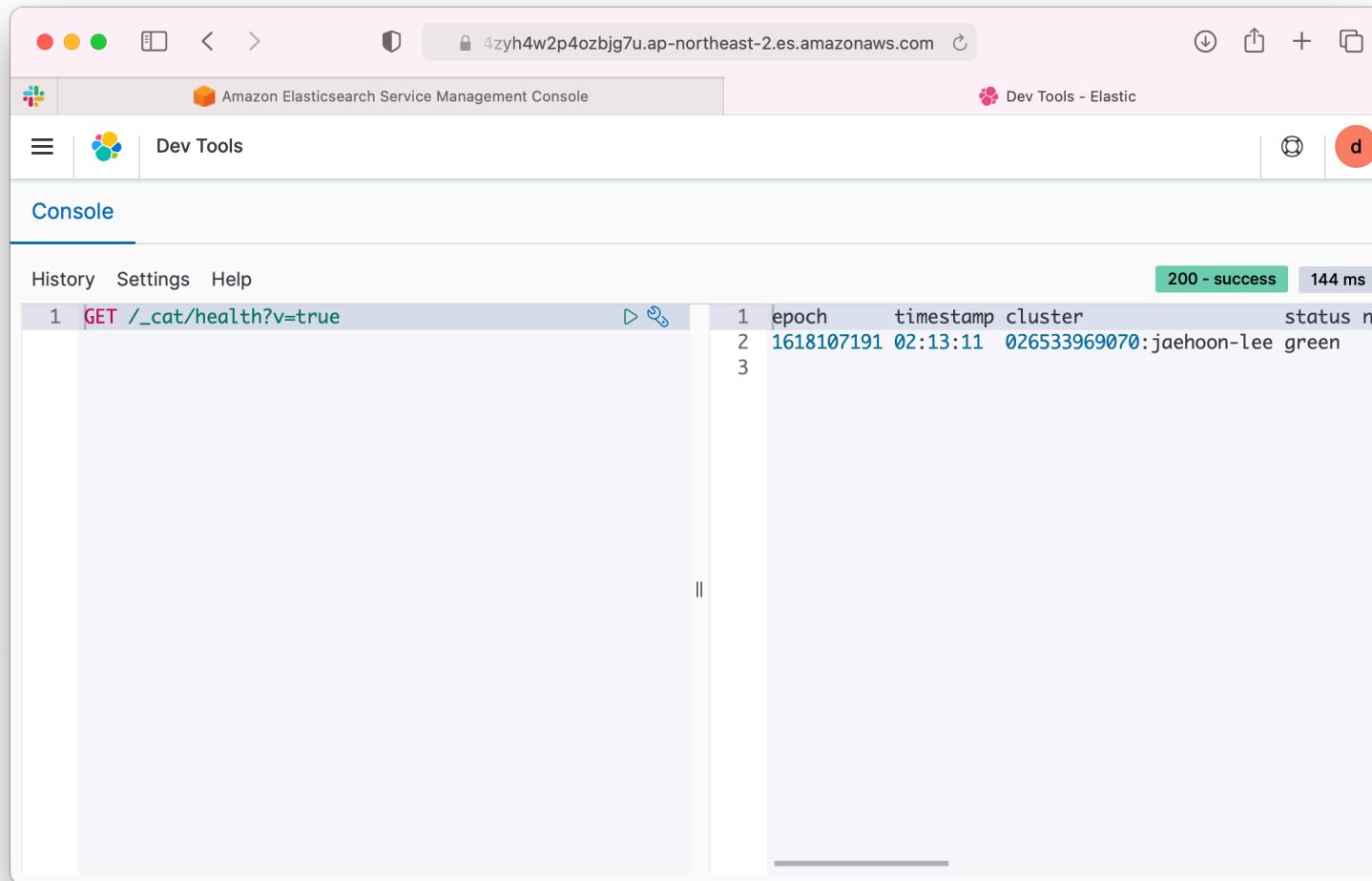


ElasticSearch Query

BAF675 금융 빅데이터 분석

이재훈, Week 2

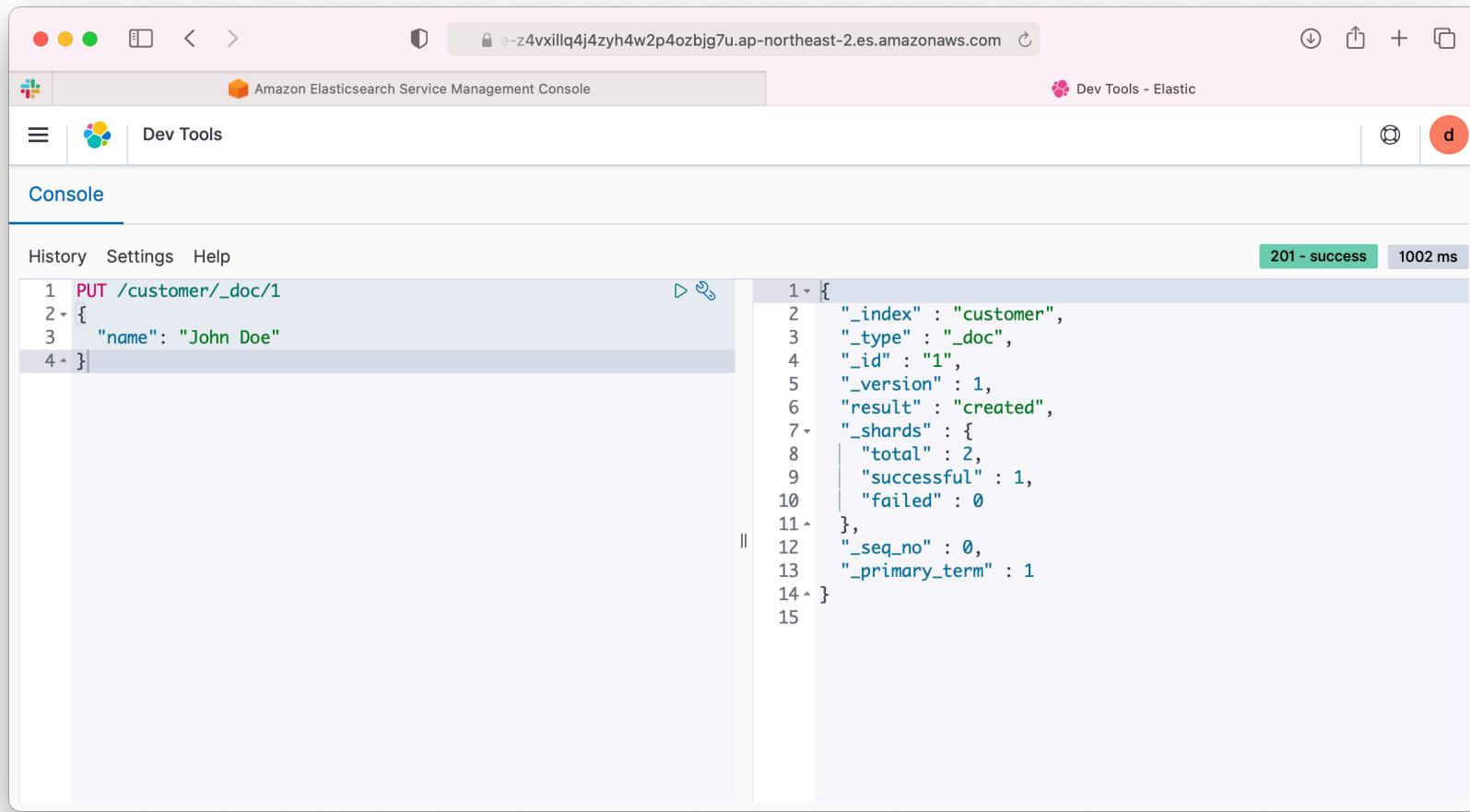
Kibana console 접속



HTTP request methods

- GET
- PUT / POST
- DELETE

새로운 doc 입력



The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools interface. The URL in the address bar is `e-z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The main area is titled "Console". In the left panel, there is a code editor with the following JSON input:

```
1 PUT /customer/_doc/1
2 {
3   "name": "John Doe"
4 }
```

The right panel displays the response from the server:

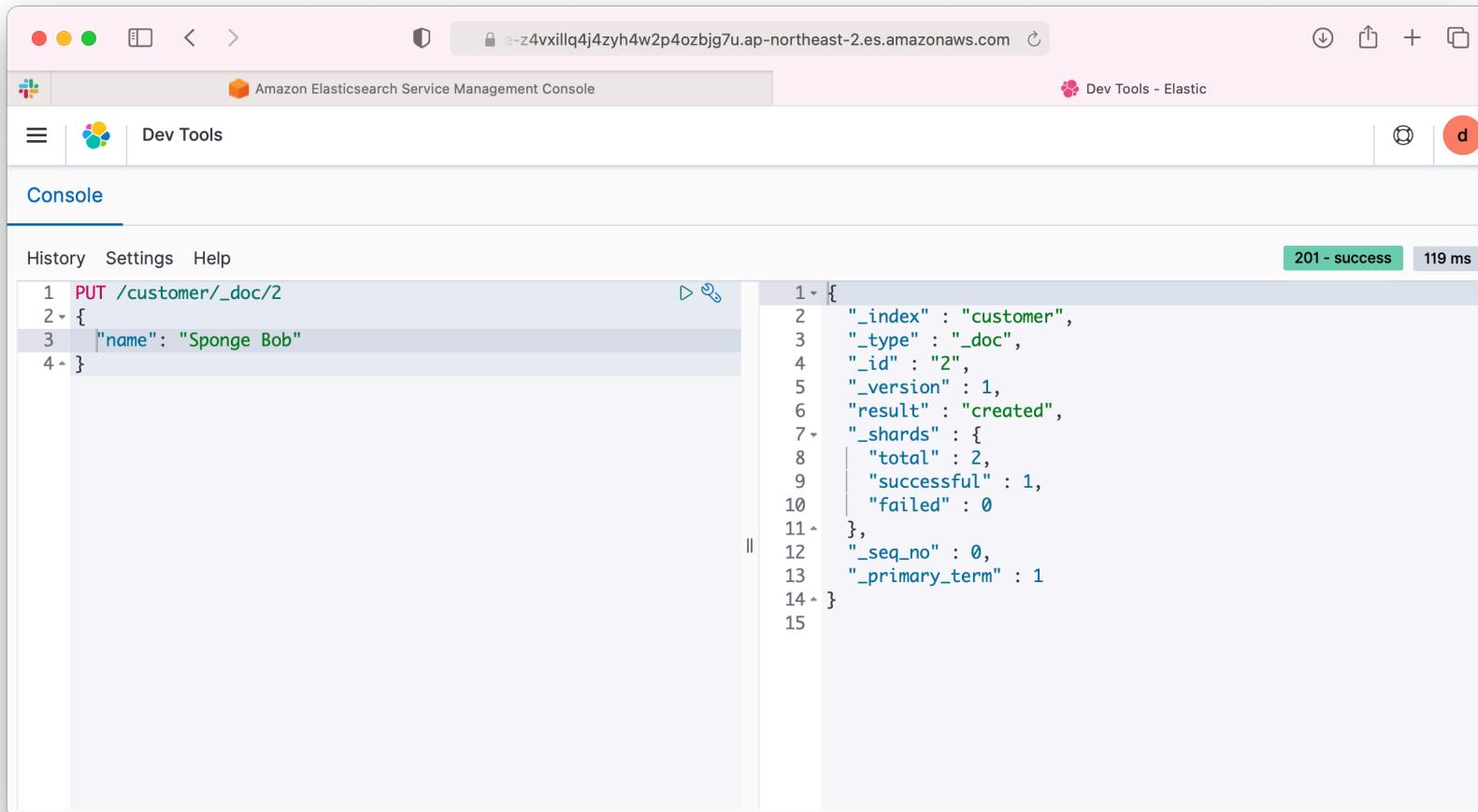
```
1 {
2   "_index" : "customer",
3   "_type" : "_doc",
4   "_id" : "1",
5   "_version" : 1,
6   "result" : "created",
7   "_shards" : {
8     "total" : 2,
9     "successful" : 1,
10    "failed" : 0
11  },
12  "_seq_no" : 0,
13  "_primary_term" : 1
14 }
15
```

The status bar at the bottom right indicates "201 - success" and "1002 ms".

쿼리 해석

- SQL 이 row 단위로 데이터를 관리한다면, ElasticSearch 는 JSON document (doc) 단위로 기록 관리
- customer 라는 이름의 새로운 index 생성
 - ElasticSearch 의 index 는 SQL 의 table 같은 개념
- Mapping concepts across SQL and Elasticsearch
 - <https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping-concepts-across-sql-and-elasticsearch.html>
- customer 라는 index 내 id = 1 로 John Doe 라는 고객 정보 doc 생성

Let's add one more



The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools - Elastic interface. The URL in the address bar is `e-z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The main area is titled "Console". In the left panel, there is a code editor with the following JSON input:

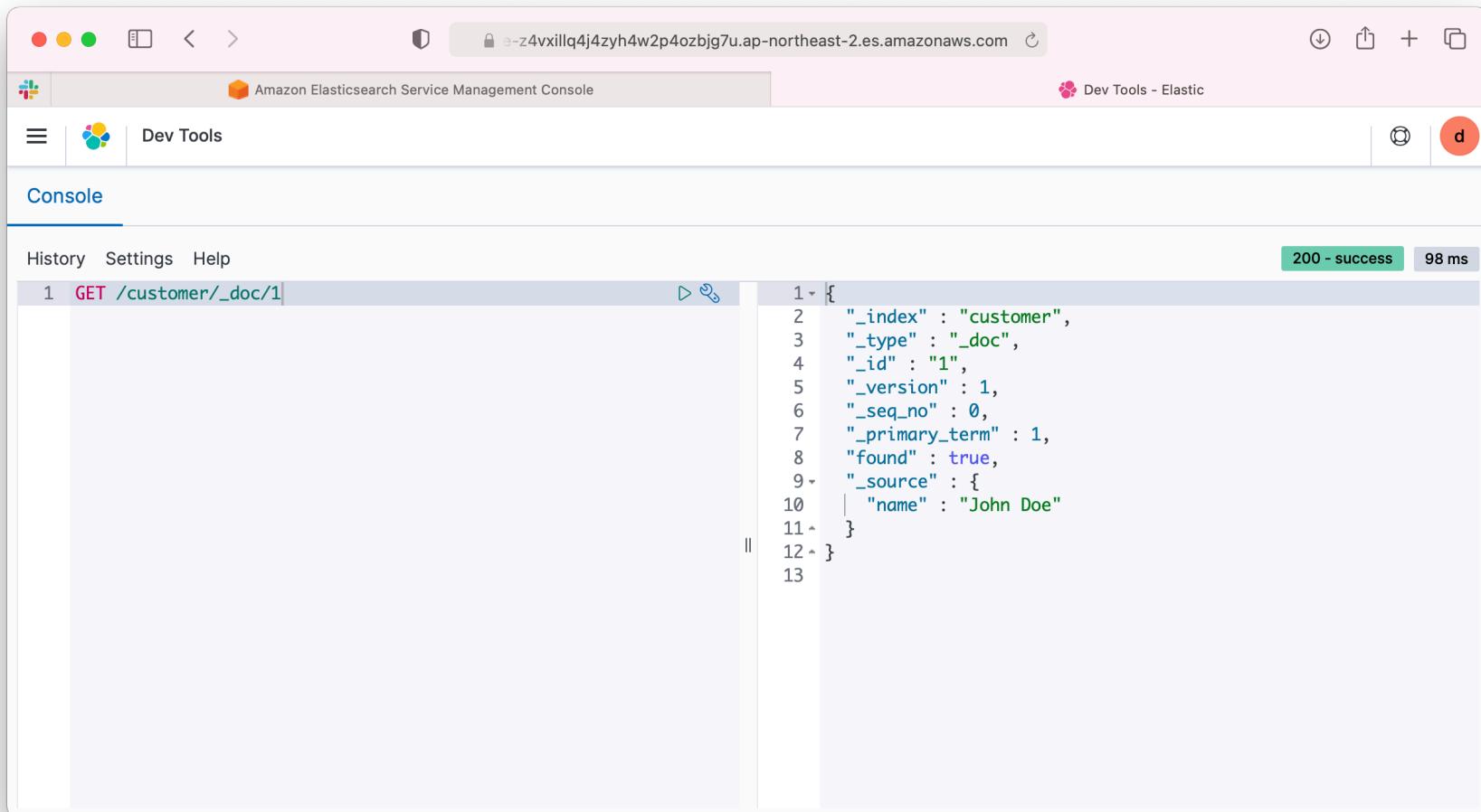
```
1 PUT /customer/_doc/2
2 {
3   "name": "Sponge Bob"
4 }
```

The response on the right is a JSON object with the following structure:

```
1 {
2   "_index" : "customer",
3   "_type" : "_doc",
4   "_id" : "2",
5   "_version" : 1,
6   "result" : "created",
7   "_shards" : {
8     "total" : 2,
9     "successful" : 1,
10    "failed" : 0
11  },
12  "_seq_no" : 0,
13  "_primary_term" : 1
14}
15
```

The status bar at the bottom right indicates "201 - success" and "119 ms".

방금 입력한 데이터 조회

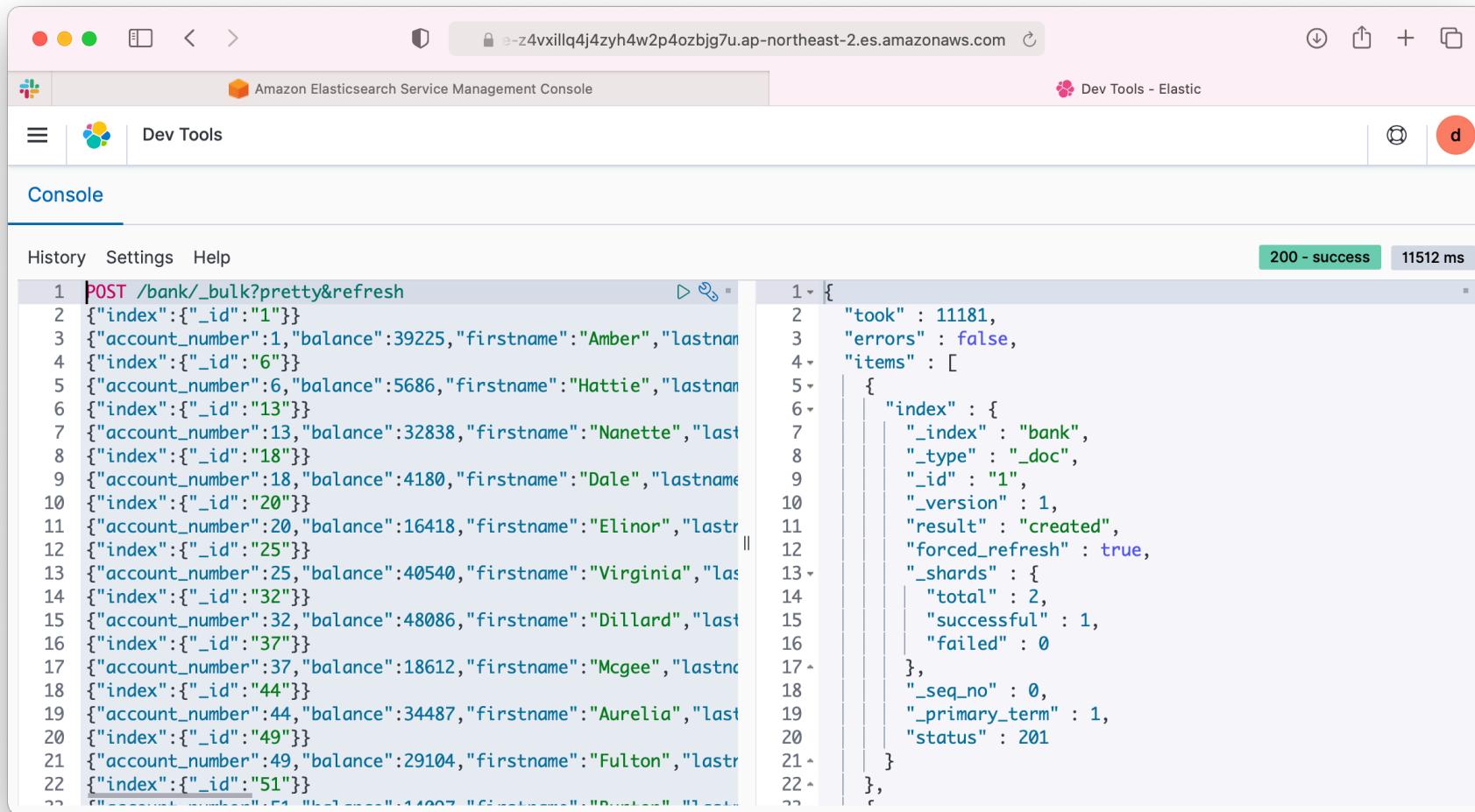


The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools - Elastic Dev Tools interface. A successful GET request is displayed:

```
1 GET /customer/_doc/1
2
3 {
4   "_index": "customer",
5   "_type": "_doc",
6   "_id": "1",
7   "_version": 1,
8   "_seq_no": 0,
9   "_primary_term": 1,
10  "found": true,
11  "_source": {
12    "name": "John Doe"
13  }
14}
```

The response status is 200 - success with a duration of 98 ms.

Bulk 데이터 입력

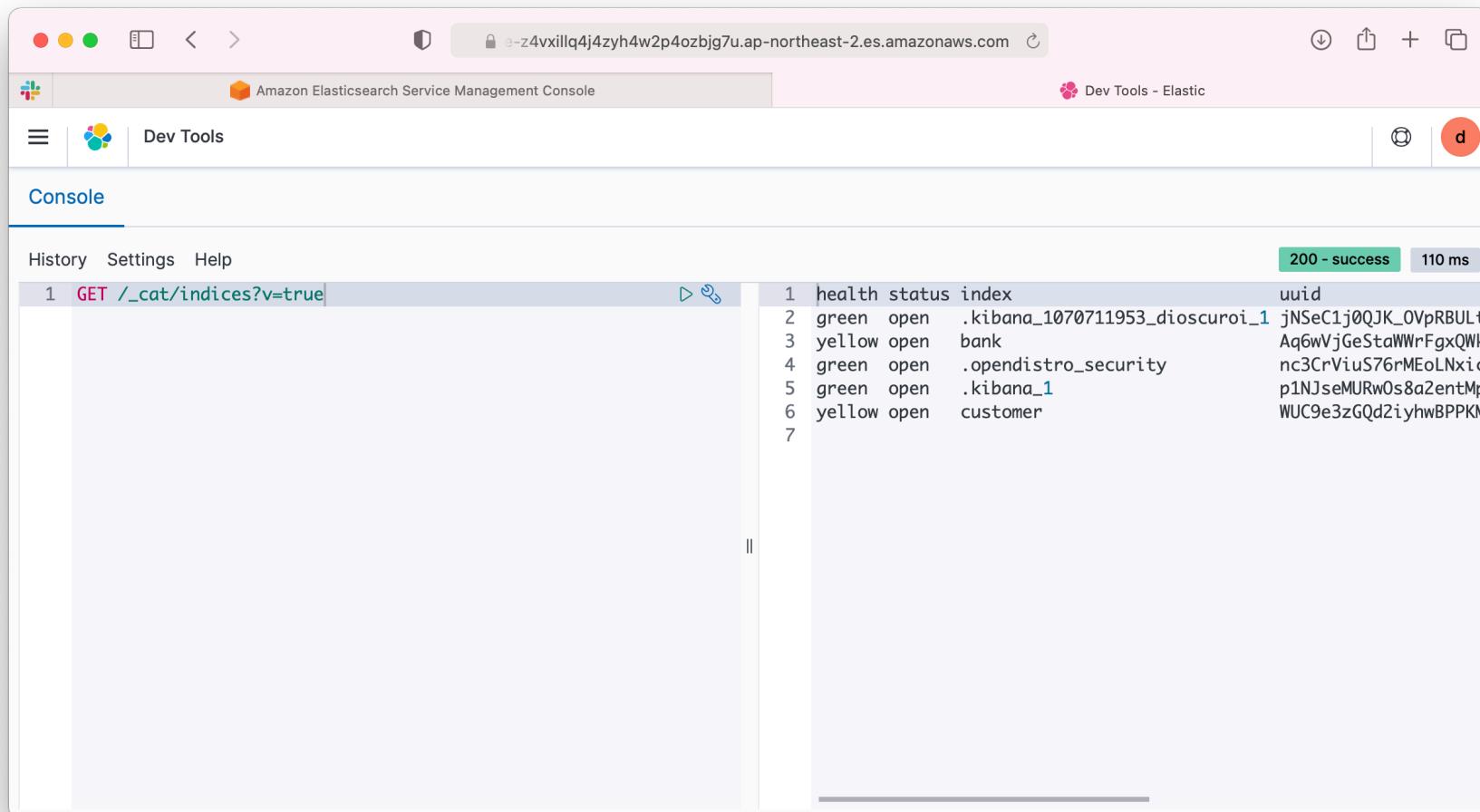


The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools - Elastic interface. A successful bulk operation has been performed on the 'bank' index. The response details the process: took 11181 milliseconds, errors false, and one item successfully created with _id 1, _version 1, result 'created', forced refresh true, and shards total 2, successful 1, failed 0. The '_seq_no' is 0, '_primary_term' is 1, and the status is 201.

```
POST /bank/_bulk?pretty&refresh
{"index":{"_id":"1"}}
{"account_number":1,"balance":39225,"firstname":"Amber","lastname":"Bartell"}
{"index":{"_id":"6"}}
{"account_number":6,"balance":5686,"firstname":"Hattie","lastname":"Casper"}
{"index":{"_id":"13"}}
 {"account_number":13,"balance":32838,"firstname":"Nanette","lastname":"Dietrich"}
 {"index":{"_id":"18"}}
 {"account_number":18,"balance":4180,"firstname":"Dale","lastname":"Fischer"}
 {"index":{"_id":"20"}}
 {"account_number":20,"balance":16418,"firstname":"Elinor","lastname":"Gandy"}
 {"index":{"_id":"25"}}
 {"account_number":25,"balance":40540,"firstname":"Virginia","lastname":"Hartmann"}
 {"index":{"_id":"32"}}
 {"account_number":32,"balance":48086,"firstname":"Dillard","lastname":"Hartmann"}
 {"index":{"_id":"37"}}
 {"account_number":37,"balance":18612,"firstname":"Mcgee","lastname":"Hartmann"}
 {"index":{"_id":"44"}}
 {"account_number":44,"balance":34487,"firstname":"Aurelia","lastname":"Hartmann"}
 {"index":{"_id":"49"}}
 {"account_number":49,"balance":29104,"firstname":"Fulton","lastname":"Hartmann"}
 {"index":{"_id":"51"}}

{
  "took" : 11181,
  "errors" : false,
  "items" : [
    {
      "index" : {
        "_index" : "bank",
        "_type" : "_doc",
        "_id" : "1",
        "_version" : 1,
        "result" : "created",
        "forced_refresh" : true,
        "shards" : {
          "total" : 2,
          "successful" : 1,
          "failed" : 0
        },
        "_seq_no" : 0,
        "_primary_term" : 1,
        "status" : 201
      }
    ]
  }
}
```

데이터 입력 상황 확인



The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools - Elastic interface. The URL in the address bar is `https://z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The main area displays the results of a `GET /_cat/indices?v=true` request. The response is a table with the following data:

index	status	index	uuid
health	status	index	
green	open	.kibana_1070711953_dioscuri_1	jNSeC1j0QJK_0VpRBUL1
yellow	open	bank	Aq6wVjGeStaWWrFgxQW1
green	open	.opendistro_security	nc3CrViuS76rMEoLNxi
green	open	.kibana_1	p1NJseMURwOs8a2entMp
yellow	open	customer	WUC9e3zGQd2iyhwBPPKM
7			

Index status

- Health
 - Green: all good
 - Yellow: primary shards are allocated but replicas are not ready
 - Red: something has gone wrong.. some primary shards are not ready

검색하기

The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools interface. The URL in the address bar is `e-z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The main area is titled "Console". The request being viewed is a GET request to `/bank/_search` with the following JSON body:

```
1 GET /bank/_search
2 {
3   "query": { "match_all": {} },
4   "sort": [
5     { "account_number": "asc" }
6   ]
7 }
```

The response status is **200 - success** and it took **162 ms**. The response body is as follows:

```
1 {
2   "took" : 83,
3   "timed_out" : false,
4   "_shards" : {
5     "total" : 5,
6     "successful" : 5,
7     "skipped" : 0,
8     "failed" : 0
9   },
10  "hits" : {
11    "total" : {
12      "value" : 1000,
13      "relation" : "eq"
14    },
15    "max_score" : null,
16    "hits" : [
17      {
18        "_index" : "bank",
19        "_type" : "_doc",
20        "_id" : "0",
21        "_score" : null,
22        "_source" : {
23          "-----": "-----"
24        }
25      }
26    ]
27  }
28}
```

검색 결과

- took: 검색에 걸린 시간
- hits: 검색 결과 포함
 - 검색 결과 전체 갯수는 1,000 개이지만 그 중 상위 10개만 리턴

다음 10개 요청하기

The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools - Elastic interface. The URL in the address bar is `e-z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The search query in the Dev Tools tab is:

```
1 GET /bank/_search
2 {
3   "query": { "match_all": {} },
4   "sort": [
5     { "account_number": "asc" }
6   ],
7   "from": 10,
8   "size": 10
9 }
```

The response is:

```
1 {
2   "took": 165,
3   "timed_out": false,
4   "_shards": {
5     "total": 5,
6     "successful": 5,
7     "skipped": 0,
8     "failed": 0
9   },
10  "hits": {
11    "total": {
12      "value": 1000,
13      "relation": "eq"
14    },
15    "max_score": null,
16    "hits": [
17      {
18        "_index": "bank",
19        "_type": "_doc",
20        "_id": "10",
21        "_score": null,
22        "_source": {
23          "-----": 10
24        }
25      }
26    ]
27  }
28}
```

A red oval highlights the `"from": 10,` and `"size": 10` parameters in the search query.

검색: 필드 지정

The screenshot shows the Elasticsearch Dev Tools Console interface. The URL in the address bar is `z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The response status is `200 - success` with a `920 ms` latency.

The search query in the console is:

```
1 GET /bank/_search
2 {
3   "query": {
4     "match": {
5       "address": "mill lane"
6     }
7   }
8 }
```

The search results are:

```
11 {
12   "total" : {
13     "value" : 19,
14     "relation" : "eq"
15   },
16   "max_score" : 7.744212,
17   "hits" : [
18     {
19       "_index" : "bank",
20       "_type" : "_doc",
21       "_id" : "136",
22       "_score" : 7.744212,
23       "_source" : {
24         "account_number" : 136,
25         "balance" : 45801,
26         "firstname" : "Winnie",
27         "lastname" : "Holland",
28         "age" : 38,
29         "gender" : "M",
30         "address" : "198 Mill Lane",
31         "employer" : "Neteria",
32         "email" : "winnieholland@neteria.com",
33         "city" : "Inie"
```

Three specific fields in the results are circled in red: `max_score`, `_id`, and `address`.

검색: 개별 단어 대신 phrase 비교

The screenshot shows the Elasticsearch Dev Tools Console interface. The URL bar indicates the request is directed to `z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The main area displays a search query and its corresponding JSON response.

Query (Left):

```
1 GET /bank/_search
2 {
3   "query": {
4     "match_phrase": {
5       "address": "mill lane"
6     }
7   }
8 }
```

Response (Right):

```
7   "skipped" : 0,
8   "failed" : 0
9 },
10 {
11   "hits" : [
12     {
13       "total" : 1,
14       "value" : 1,
15       "relation" : "eq"
16     }
17   ],
18   "max_score" : 7.744211,
19   "hits" : [
20     {
21       "_index" : "bank",
22       "_type" : "_doc",
23       "_id" : "136",
24       "_score" : 7.744211,
25       "_source" : {
26         "account_number" : 136,
27         "balance" : 45801,
28         "firstname" : "Winnie",
29         "lastname" : "Holland",
30         "age" : 38,
31         "gender" : "M"
32     }
33   }
34 }
```

The JSON response is annotated with red circles highlighting the `"hits"` array at line 10 and the `"value"` field at line 12, indicating the result count.

검색: bool 로 여러 조건 조합하기

- must, should, must_not 으로 조합

The screenshot shows the Elasticsearch Dev Tools Console interface. The URL bar indicates the connection is to an AWS Elasticsearch cluster at `z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The console tab is active, and the history, settings, and help tabs are visible at the top.

The query entered in the console is:

```
1 GET /bank/_search
2 {
3   "query": {
4     "bool": {
5       "must": [
6         { "match": { "age": "40" } }
7       ],
8       "must_not": [
9         { "match": { "state": "ID" } }
10      ]
11    }
12  }
13 }
```

The response returned is:

```
1 {
2   "took" : 80,
3   "timed_out" : false,
4   "_shards" : {
5     "total" : 5,
6     "successful" : 5,
7     "skipped" : 0,
8     "failed" : 0
9   },
10  "hits" : {
11    "total" : {
12      "value" : 43,
13      "relation" : "eq"
14    },
15    "max_score" : 1.0,
16    "hits" : [
17      {
18        "_index" : "bank",
19        "_type" : "_doc",
20        "_id" : "177",
21        "_score" : 1.0,
22        "._source" : {
23          "age" : "40",
24          "name" : "John Doe",
25          "state" : "CA",
26          "balance" : 1000000
27        }
28      }
29    ]
30  }
31}
```

The response status is 200 - success and it took 107 ms.

검색: filter로 검색 결과 제한하기

- 다른 query 와의 차이점: filter 는 score 에 영향을 주지 않는다

The screenshot shows the Elasticsearch Dev Tools Console interface. The URL in the address bar is `z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The console tab is selected, showing a search query and its results.

History Settings Help 200 - success 112 ms

```
1 GET /bank/_search
2 {
3   "query": {
4     "bool": {
5       "must": { "match_all": {} },
6       "filter": {
7         "range": {
8           "balance": {
9             "gte": 20000,
10            "lte": 30000
11          }
12        }
13      }
14    }
15  }
16 }
```

The results section shows the response from the search query. It includes the total number of hits (217), the maximum score (1.0), and the first hit's details:

```
10  "hits" : {
11    "total" : {
12      "value" : 217,
13      "relation" : "eq"
14    },
15    "max_score" : 1.0,
16    "hits" : [
17      {
18        "_index" : "bank",
19        "_type" : "_doc",
20        "_id" : "49",
21        "score" : 1.0
22      }
23    ]
24  }
25  "source" : {
26    "account_number" : 49,
27    "balance" : 29104,
28    "firstname" : "Fulton",
29    "lastname" : "Holt",
30    "age" : 23,
    "gender" : "F",
    "address" : "451 Humboldt Street",
    "employer" : "Anocha",
  }
```

Two specific fields in the result object are circled in red: `max_score` and `score`. These are highlighted because they are typically affected by the `query` part of the search, but here they are shown as 1.0, indicating that the `filter` did not impact the scoring of the results.

Aggregations

- pandas 에서 하던 group-by 기능
 - 분류별 갯수 세기, 합계, 평균 등

Aggregations: count

```
1 GET /bank/_search
2 {
3   "size": 0,
4   "aggs": {
5     "group_by_state": {
6       "terms": {
7         "field": "state.keyword"
8       }
9     }
10  }
11 }
```

```
15   "max_score" : null,
16   "hits" : [ ]
17 },
18 },
19 "group_by_state" : [
20   "doc_count_error_upper_bound" : 18,
21   "sum_other_doc_count" : 764,
22   "buckets" : [
23     {
24       "key" : "TX",
25       "doc_count" : 30
26     },
27     {
28       "key" : "ID",
29       "doc_count" : 27
30     },
31     {
32       "key" : "MD",
33       "doc_count" : 25
34     },
35     {
36       "key" : "MA"
37     }
38   ]
39 }
```

Aggregations: count - 52개 주 모두 요청하기

The screenshot shows the Elasticsearch Dev Tools Console interface. The top bar includes standard OS X window controls, a URL bar with the address 'jon-lee-z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com', and a tab labeled 'Dev Tools - Elastic'. Below the tabs are icons for 'Dashboard' and 'Visualize'. The main area is titled 'Console'.

The 'History' tab is selected, showing a single search request:

```
1 GET /bank/_search
2 {
3   "size": 0,
4   "aggs": {
5     "group_by_state_gender": {
6       "terms": {
7         "field": "state.keyword",
8         "size": 100
9       }
10      }
11    }
12 }
```

The response is displayed in the right pane, showing a list of states and their document counts:

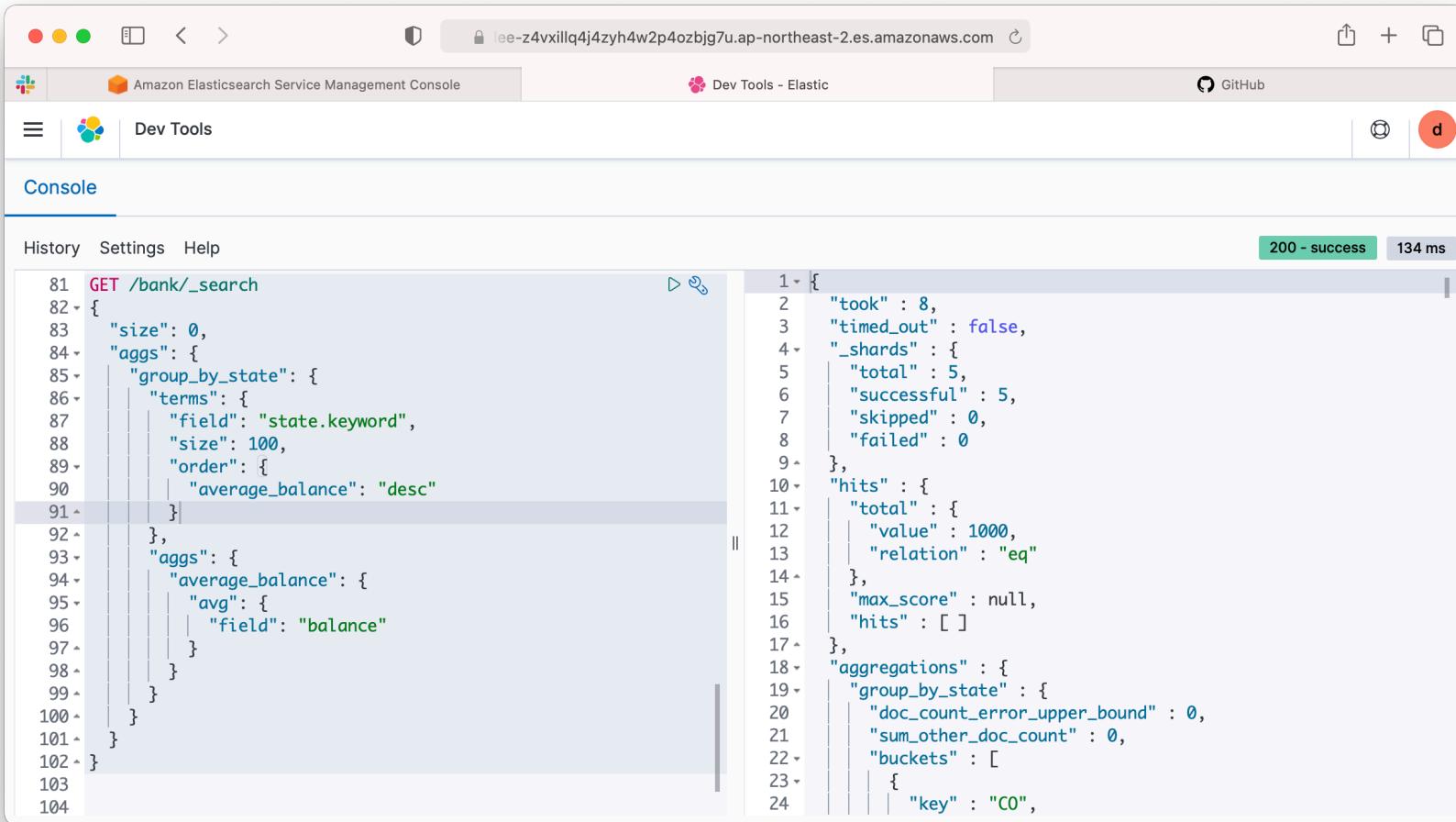
```
211 [
212   {
213     "key" : "DE",
214     "doc_count" : 14
215   },
216   {
217     "key" : "NM",
218     "doc_count" : 14
219   },
220   {
221     "key" : "NV",
222     "doc_count" : 13
223   },
224   {
225     "key" : "SC",
226     "doc_count" : 13
227   }
228 ]
229
230 }
```

The status bar at the bottom indicates '200 - success' and '53 ms'.

Aggregations: average balance

```
1 GET /bank/_search
2 {
3   "size": 0,
4   "aggs": {
5     "group_by_state": {
6       "terms": {
7         "field": "state.keyword"
8       },
9       "aggs": {
10      "average_balance": {
11        "avg": {
12          "field": "balance"
13        }
14      }
15    }
16  }
17 }
18 }
19 {
20   "aggregations": {
21     "group_by_state": {
22       "doc_count_error_upper_bound": 18,
23       "sum_other_doc_count": 764,
24       "buckets": [
25         {
26           "key": "TX",
27           "doc_count": 30,
28           "average_balance": {
29             "value": 26073.3
30           }
31         },
32         {
33           "key": "ID",
34           "doc_count": 27,
35           "average_balance": {
36             "value": 24368.777777777777
37           }
38         },
39         {
40           "key": "MD"
41         }
42       ]
43     }
44   }
45 }
```

Aggregations: sort by average balance



The screenshot shows the Amazon Elasticsearch Service Management Console Dev Tools tab. The URL is `lee-z4vxillq4j4zyh4w2p4ozbjg7u.ap-northeast-2.es.amazonaws.com`. The response status is `200 - success` with a `134 ms` latency.

```
81 GET /bank/_search
82 {
83   "size": 0,
84   "aggs": {
85     "group_by_state": {
86       "terms": {
87         "field": "state.keyword",
88         "size": 100,
89         "order": {
90           "average_balance": "desc"
91         }
92       },
93       "aggs": {
94         "average_balance": {
95           "avg": {
96             "field": "balance"
97           }
98         }
99       }
100     }
101   }
102 }
```

```
1 {
2   "took": 8,
3   "timed_out": false,
4   "_shards": {
5     "total": 5,
6     "successful": 5,
7     "skipped": 0,
8     "failed": 0
9   },
10  "hits": {
11    "total": {
12      "value": 1000,
13      "relation": "eq"
14    },
15    "max_score": null,
16    "hits": []
17  },
18  "aggregations": {
19    "group_by_state": {
20      "doc_count_error_upper_bound": 0,
21      "sum_other_doc_count": 0,
22      "buckets": [
23        {
24          "key": "CO",
```