Labo Examen Big Data

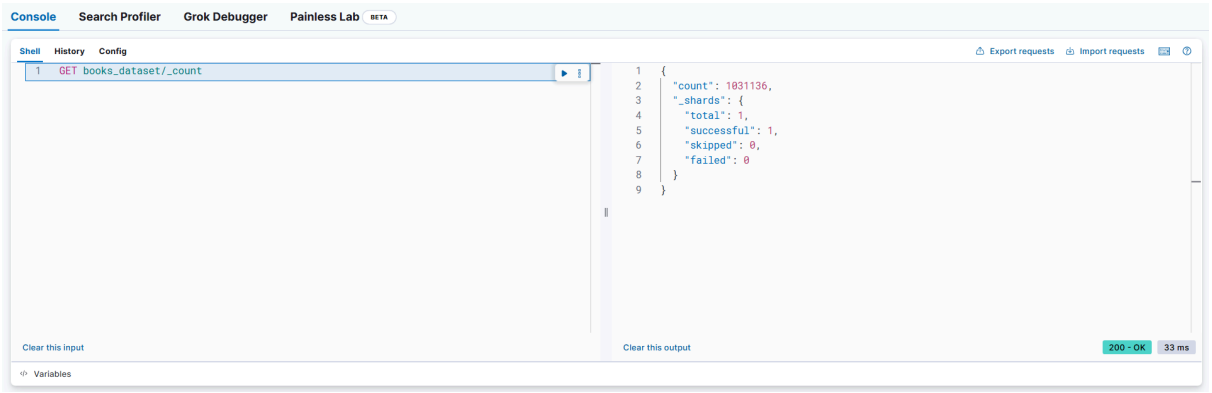
Queries & aggregaties

De heel eenvoudige

Hoeveel documenten zijn er in totaal?

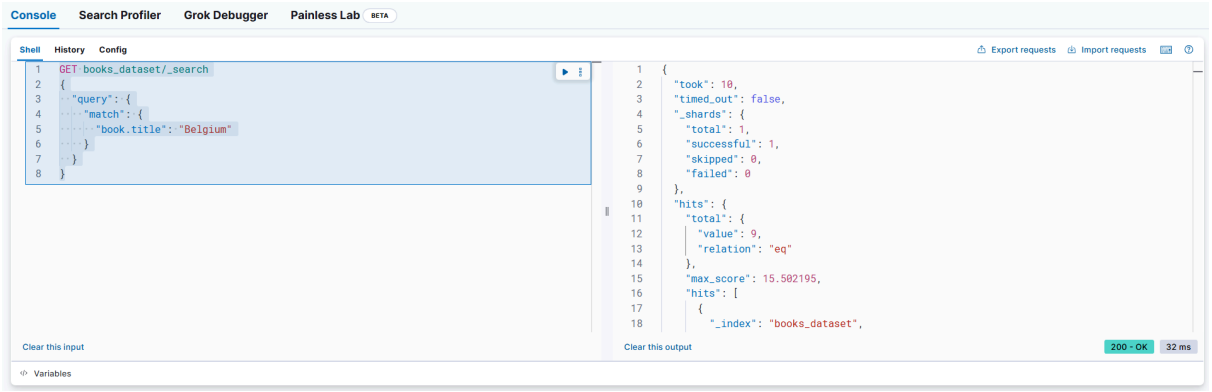
1031136

```
GET books_dataset/_count
```



Zoek alle documenten waarvan de titel “Belgium” bevat.

```
GET books_dataset/_search
{
  "query": {
    "match": {
      "book.title": "Belgium"
    }
  }
}
```



De eenvoudige

Wat is de gemiddelde score van een boek?

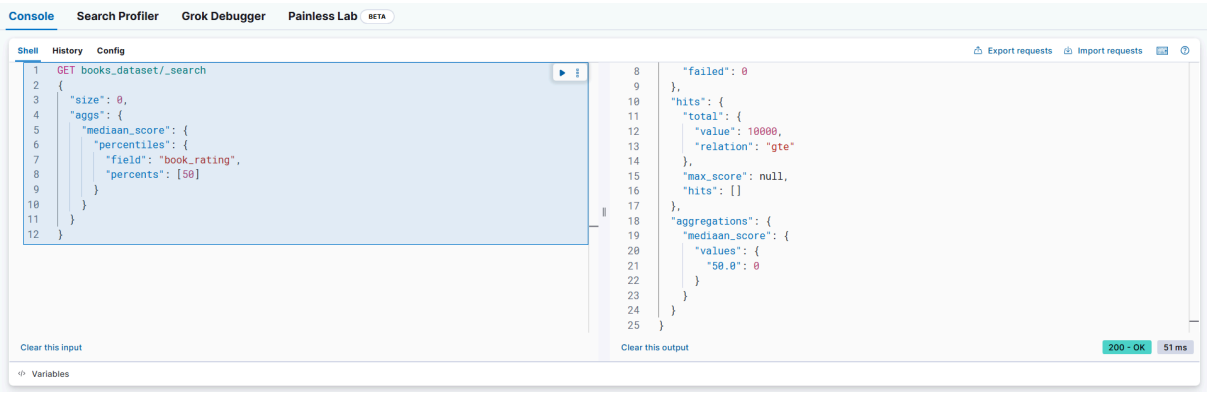
2.84

```
GET books_dataset/_search
{
  "size": 0,
  "aggs": {
    "gemiddelde_score": {
      "avg": {
        "field": "book_rating"
      }
    }
  }
}
```

Wat is de mediaanscore van een boek?

0

```
GET books_dataset/_search
{
  "size": 0,
  "aggs": {
    "mediaan_score": {
      "percentiles": {
        "field": "book_rating",
        "percents": [50]
      }
    }
  }
}
```



Iets moeilijker

Welke gebruiker (user_id) heeft in totaal de meeste scores gepost? Uit welk land komt deze gebruiker? Probeer overbodigevelden uit het resultaat te filteren.

11676

```
GET books_dataset/_search
{
  "size": 1,
  "aggs": {
    "meeste_scores": {
      "terms": {
        "field": "user.user_id",
```

```
      "size": 1,
      "order": {
       "_count": "desc"
      }
     },
     "aggs": {
      "land": {
       "top_hits": {
        "_source": ["user.user_id", "user.country"],
        "size": 1
       }
      }
     }
    }
  }
}
```

Moeilijker

Welk boek heeft de hoogste gemiddelde score binnen de 10 meest beoordeelde boeken?

Harry Potter and the Chamber of Secrets (Book 2)

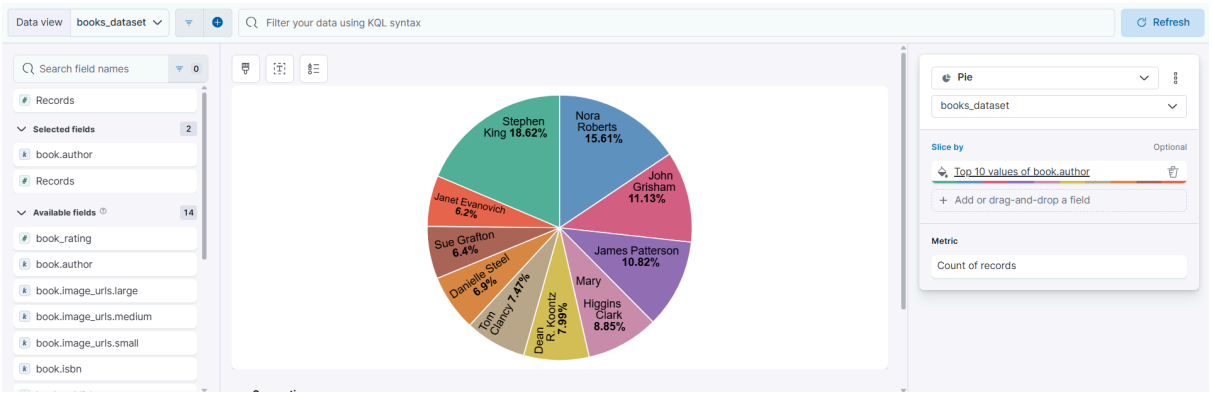
```
GET books_dataset/_search
{
 "size": 0,
 "aggs": {
  "top_10_meest_beoordeeld": {
```

```
      "terms": {
        "field": "book.isbn",
        "size": 1000,
        "order": {
          "aantal_ratings": "desc"
        }
      },
      "aggs": {
        "aantal_ratings": {
          "value_count": {
            "field": "book_rating"
          }
        },
        "gemiddelde_score": {
          "avg": {
            "field": "book_rating"
          }
        },
        "boek_info": {
          "top_hits": {
            "_source": ["book.title", "book.author", "book.isbn"],
            "size": 1
          }
        },
        "sort_op_gemiddelde": {
          "bucket_sort": {
            "sort": [
              { "gemiddelde_score": { "order": "desc" } }
            ],
            "size": 1
          }
        }
      }
    }
  }
}
```

Visualisatie

Probeer de onderstaande visualisatie na te maken. Wat toont deze eigenlijk?

De taartgrafiek toont de 10 auteurs met de meeste boeken/beoordelingen in de dataset, elk als een segment van het totaal.



Clusterstatus

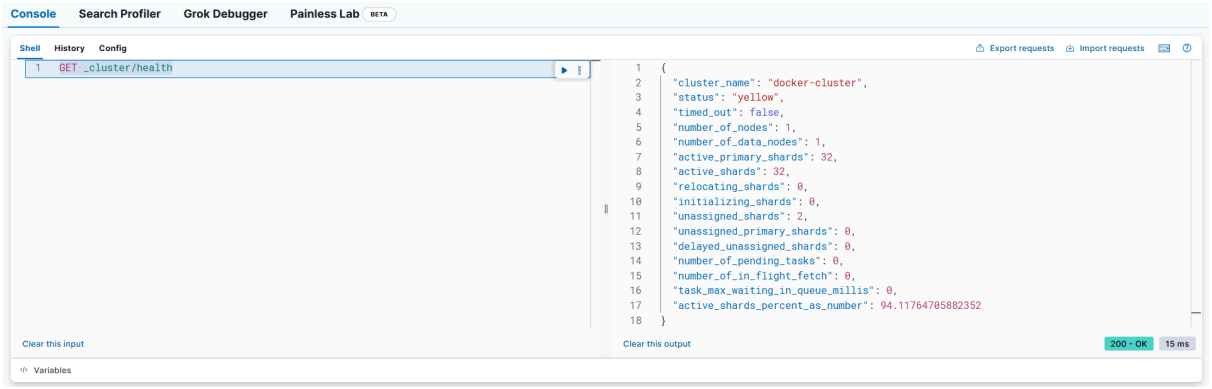
Wat is de status van de cluster? Leg uit!

De status van de cluster geeft aan of Elasticsearch correct functioneert:

- **green = Alles werkt perfect**
- **yellow = Data is beschikbaar, maar replica's ontbreken**
- **red = Sommige data is onbeschikbaar of corrupt**

Je checkt het met:

```
GET _cluster/health
```


Console
Search Profiler
Grok Debugger
Painless Lab
BETA
Shell
History
Config
Export requests
Import requests
1 GET _cluster/health
Clear this input
Clear this output
200 - OK
15 ms
Variables
{
"cluster_name": "docker-cluster",
"status": "yellow",
"timed_out": false,
"number_of_nodes": 1,
"number_of_data_nodes": 1,
"active_primary_shards": 32,
"active_shards": 32,
"relocating_shards": 0,
"initializing_shards": 0,
"unassigned_shards": 2,
"unassigned_primary_shards": 0,
"delayed_unassigned_shards": 0,
"number_of_pending_tasks": 0,
"number_of_in_flight_fetch": 0,
"task_max_waiting_in_queue_millis": 0,
"active_shards_percent_as_number": 94.11764705882352
}