13 - Aggregations

OVERVIEW

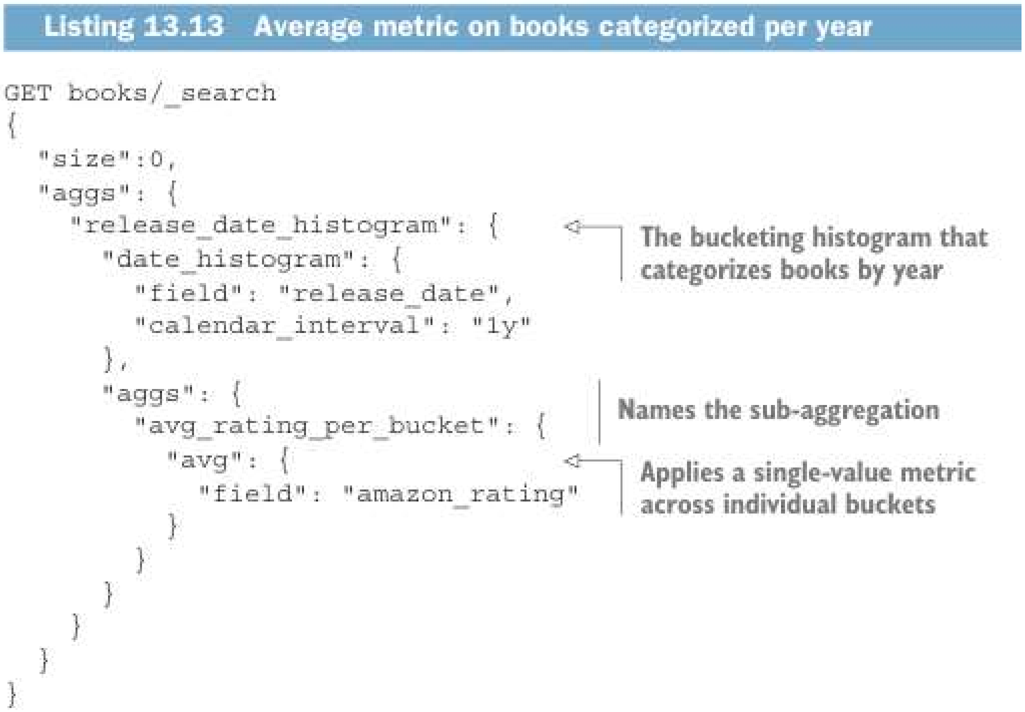
- Metric aggregations
  - métricas tipo soma ou média
- Bucket aggregations
  - Juntar dados em intervalos
- Pipeline aggregations
  - Linka aggregations anteriores para aggregations complexos
- Usa o mesmo _search endpoint das queries comuns, mas com aggs no body
- Podemos combinar query com aggregations, gera as aggregations sobre o resultado da query.
- Podemos fazer várias aggregations ao mesmo tempo e aninhadas
- Setamos size=0 na query para não retornar os documentos em si, só as agregações.

METRIC AGGREGATIONS

- Somas, médias, mínimo, etc.
- value_count: contagem simples sobre uma variável booleana
- Aggregations não são feitas para text fields:
  - Causa erro
  - Pode ser contornado alterando o parâmetro fielddata no mapping
  - Não recomendado, desempenho ruim
- stats retorna várias estatísticas de uma vez (ou extended_stats para mais ainda)
- cardinality retorna número de valores únicos de um campo.

BUCKET AGGREGATIONS

- Agrupa documentos segundo alguma critério
- e.g. para criar histogramas
  - Definimos os bins manualmente (ranges) na query ou por intervalo (interval)
    - Podemos agregar IPs com o parâmetro "ip_range"
  - Também pode ser por intervalo de data (date_histogram)
    - Pode ser calendar-aware (calendar_interval) ou não (fixed_interval)
- Podemos ter buckets dentro de buckets (sub-aggregation):



Listing 13.13 Average metric on books categorized per year

```
GET books/_search
{
  "size":0,
  "aggs": {
    "release_date_histogram": {
      "date_histogram": {
        "field": "release_date",
        "calendar_interval": "1y"
      },
      "aggs": {
        "avg_rating_per_bucket": {
          "avg": {
            "field": "amazon_rating"
          }
        }
      }
    }
  }
}
```

The bucketing histogram that categorizes books by year

Names the sub-aggregation

Applies a single-value metric across individual buckets

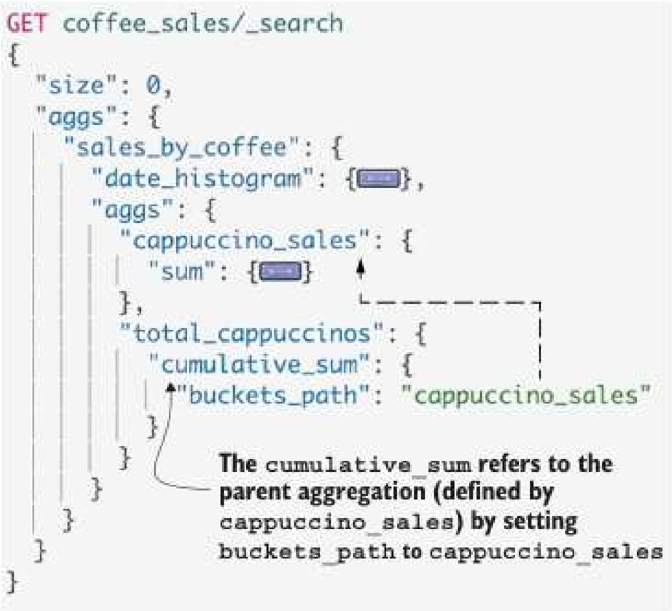
- Podemos agregar itens pela presença de um dado termo em algum campo (e.g. nome de autor)
  - Ou ainda um conjunto de condições, não apenas um único termo. (multi_terms)

PARENT AND SIBLING AGGREGATIONS

- Há dois tipos de agregations: parent e sibling
- Parent aggregations:
  - aggregation sobre uma outra agregação
    - e.g. média por dia
- Sibling aggregations:
  - agregações em paralelo, todas feitas em cima do conjunto original.
    - e.g. média e histograma

PIPELINE AGGREGATIONS

- Concatenar várias agregações
- Também tem os dois tipos sibling e parent
- Parent:



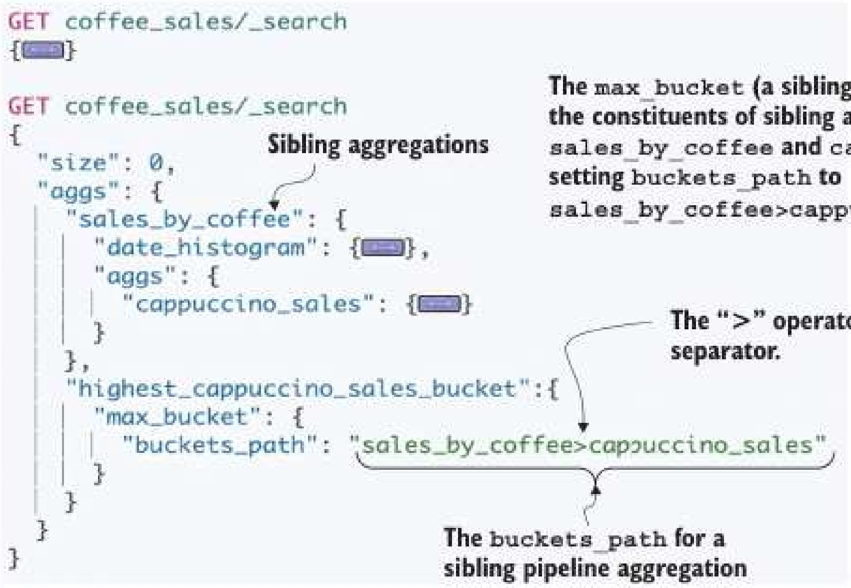
```
GET coffee_sales/_search
{
  "size": 0,
  "aggs": {
    "sales_by_coffee": {
      "date_histogram": {…},
      "aggs": {
        "cappuccino_sales": {
          "sum": {…}
        },
        "total_cappuccinos": {
          "cumulative_sum": {
            "buckets_path": "cappuccino_sales"
          }
        }
      }
    }
  }
}
```

The cumulative_sum refers to the parent aggregation (defined by cappuccino_sales) by setting buckets_path to cappuccino_sales

Figure 13.8 Parent pipeline aggregation buckets_path setting

-

  - Sibling:



```
GET coffee_sales/_search
{…}

GET coffee_sales/_search
{
  "size": 0,
  "aggs": {
    "sales_by_coffee": {
      "date_histogram": {…},
      "aggs": {
        "cappuccino_sales": {…}
      }
    },
    "highest_cappuccino_sales_bucket":{
      "max_bucket": {
        "buckets_path": "sales_by_coffee>cappuccino_sales"
      }
    }
  }
}
```

Sibling aggregations

The max_bucket (a sibling aggregation) refers to the constituents of sibling aggregations (defined by sales_by_coffee and cappuccino_sales) by setting buckets_path to sales_by_coffee>cappuccino_sales.

The ">" operator is the aggregation separator.

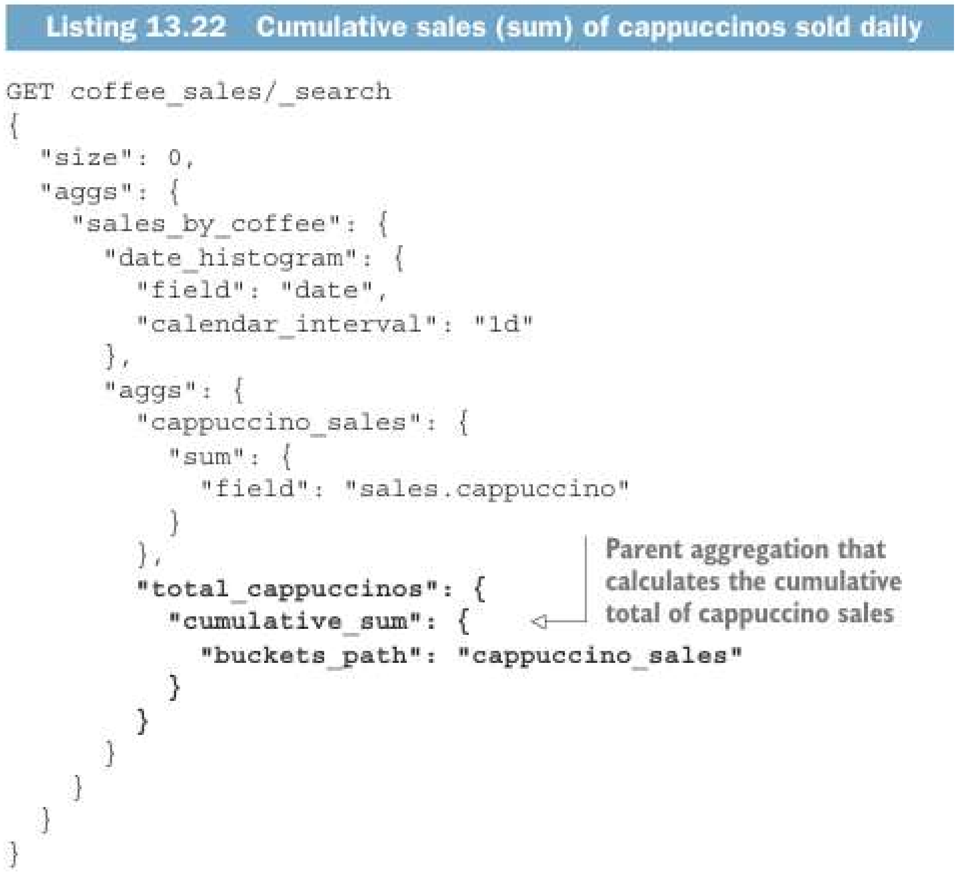
The buckets_path for a sibling pipeline aggregation

Figure 13.9 Sibling pipeline aggregation buckets_path setting

-

  - Existem várias pipeline aggregations pré definidas
- Algumas são sibling aggregations (e.g. sum_bucket), outras são parent aggregations (e.g. bucket_sort)

- cumulative_sum parent pipeline aggregation:
    - Retorna somas acumuladas



Listing 13.22 Cumulative sales (sum) of cappuccinos sold daily

```
GET coffee_sales/_search
{
  "size": 0,
  "aggs": {
    "sales_by_coffee": {
      "date_histogram": {
        "field": "date",
        "calendar_interval": "1d"
      },
      "aggs": {
        "cappuccino_sales": {
          "sum": {
            "field": "sales.cappuccino"
          }
        },
        "total_cappuccinos": {
          "cumulative_sum": {
            "buckets_path": "cappuccino_sales"
          }
        }
      }
    }
  }
}
```

Parent aggregation that calculates the cumulative total of cappuccino sales

- max_bucket e min_bucket sibling pipeline agreggation:
  - Retorna o bucket máximo/mínimo de um conjunto de buckets
  - e.g. retorna bucket (dia) que vendeu mais cappuccinos

Listing 13.23 Pipeline aggregation to find to sales of cappuccinos

```
GET coffee_sales/_search
{
  "size": 0,
  "aggs": {
    "sales_by_coffee": {
      "date_histogram": {
        "field": "date",
        "calendar_interval": "1d"
      },
      "aggs": {
        "cappuccino_sales": {
          "sum": {
            "field": "sales.cappuccino"
          }
        }
      }
    },
    "highest_cappuccino_sales_bucket":{
      "max_bucket": {
        "buckets_path": "sales_by_coffee>cappuccino_sales"
      }
    }
  }
}
```

Summary

- Whereas a search finds answers in the amassed data based on a search criterion, an aggregation compiles patterns, insights, and information for data collected by organizations.
- Elasticsearch allows us to perform nested and sibling aggregations on data.
- Elasticsearch classifies aggregations into three types: metrics, buckets, and pipelines.
- Metric aggregations fetch single-value metrics such as `avg`, `min` and `max`, `sum`, and so on.
- Bucket aggregations classify data into various buckets based on a bucketing strategy. With a bucketing strategy, we can ask Elasticsearch to split data into buckets as needed.
- We can either let Elasticsearch create predefined buckets based on the interval we provide or create custom ranges:
  - If the interval is 10 for an age group, for example, Elasticsearch splits data into steps of 10.
  - If we want to create a range like 10 to 30 or 30 to 100, where the interval differs, we can create a custom range.
- Pipeline aggregations work on the output from other metric and bucket aggregations to create new aggregations or new buckets.