Semantix

Elastic Essential I

Aula 5

Quem sou eu?

Eu sou Rodrigo Augusto Rebouças.

Engenheiro de dados da Semantix
Instrutor do Semantix Academy

Você pode me encontrar em:
rodrigo.augusto@semantix.com.br

Analyzer

Conceitos

Principais analyzer

Semantix

Analyzer Introdução

- Busca exata
  - Sim e não
- Busca FullText
  - Quanto a busca casa (_score)
  - Analisadores
    - Testar
    - Aplicar em atributos específicos
    - Analyzer personalizado
- Índice invertido
  - Quebrar em tokens
  - Inserir numa tabela
  - _search?cidade="São Paulo"
    - Tokens: são paulo

Analyzer Principais

- Espaço em branco: whitespace
  - Separa as palavras por espaço
- Simples: simple
  - Remover números
  - Remover espaços e pontuação (‘!@#$%¨&*-+´~^/:;.><,)
  - Somente texto
  - Texto em lowercase
- Padrao: standard
  - Remover espaços e pontuação
  - Texto em lowercase
- Idioma: brazilian, english
  - Remover acentos, gênero e plural

Analyzer Exemplo

- Analyzer Standard ou Simple

```
POST _analyze
{
  "analyzer": "standard",
  "text": "Elasticsearch e Hadoop são ferramentas de Big Data"
}
```

```
POST _analyze
{
  "analyzer": "simple",
  "text": "Elasticsearch e Hadoop são ferramentas de Big Data"
}
```

```
{
 "tokens" : [
  {"token" : "elasticsearch"},
  {"token" : “e"},
  {"token" : "hadoop"},
  {"token" : “são"},
  {"token" : "ferramentas"},
  {"token" : “de"},
  {"token" : "big"},
  {"token" : "data"}
 ]
}
```

Analyzer Exemplo

- Analyzer whitespace

```
POST _analyze
{
  "analyzer": "whitespace",
  "text": "Elasticsearch e Hadoop são ferramentas de Big Data"
}
```

```
{
 "tokens" : [
  {"token" : “Elasticsearch"},
  {"token" : “e"},
  {"token" : “Hadoop"},
  {"token" : “são"},
  {"token" : "ferramentas"},
  {"token" : “de"},
  {"token" : “Big"},
  {"token" : “Data"}
 ]
}
```

Analyzer Exemplo

- Analyzer em Português

```
POST _analyze
{
  "analyzer": "brazilian",
  "text": "Elasticsearch e Hadoop são ferramentas de Big Data"
}
```

```
{
 "tokens" : [
   {"token" : "elasticsearch"},
   {"token" : "hadoop"},
   {"token" : “sao"},
   {"token" : "ferrament"},
   {"token" : "big"},
   {"token" : "dat"}
 ]
}
```

Analyzer Exemplo

- Analyzer English

```
POST _analyze
{
  "analyzer": "english",
  "text": "Elasticsearch and Hadoop are Big Data tools"
}
```

```
{
 "tokens" : [
   {"token" : "elasticsearch"},
   {"token" : "hadoop"},
   {"token" : "big"},
   {"token" : "data"}
   {"token" : “tool"}
 ]
}
```

Analyzer Adicionar em um atributo

```
PUT cliente1
{
  "mappings": {
    "properties": {
      "conhecimento": {
        "type": "text",
        "analyzer": "standard"
      }
    }
  }
}
```


Semantix

Analyzer Boas Práticas

- Indexar o mesmo campo de maneiras diferentes para fins diferentes
  - Tipo Keyword
    - Classificação
    - Agregação
  - Tipo Text
    - Pesquisa Fulltext

- Manter 2 versões do atributo com analyzer
  - Tipo Keyword
    - Dado original
  - Tipo text
    - Dado com analisador

Analyzer Exemplo – Campo de 2 Tipos

```
PUT cliente2
{
  "mappings": {
    "properties": {
      "conhecimento": {
        "type": "text",
        "analyzer": "standard",
        "fields": {"raw": {"type":  "keyword"}}
      }
    }
  }
}
```

Exemplo Criação de índice com settings e mappings

```
PUT cliente3
{
 "settings": {
  "index": {
   "number_of_shards": 1,
   "number_of_replicas": 0
  }
 },
 "mappings": {
  "properties": {
   "nome": {"type": "text"},
   "conhecimento": {
    "type": "text",
    "analyzer": "whitespace",
    "fields": {
     "raw": {"type": "keyword"}
    }
   }
  }
 }
}
```

Exercícios Analyzer

1. Criar os Analyzer simple, standard, brazilian e portuguese para a seguinte frase:
   - O elasticsearch surgiu em 2010

2. Realizar os passos no índice produto
   a) Criar um analyzer brazilian para o atributo descricao
   b) Para o atributo descricao aplicar o analzyer brazilian para o tipo de campo text e criar o atributo descricao.original com o dado do tipo keyword
   c) Buscar a palavra “compativel” no campo descricao.original (hits = 0)
   d) Buscar a palavra “compativel” no campo descricao

Aggregations

Conceitos

Tipos

Semantix

Agregações Conceitos

- Forma de analisar os dados indexados
- Estrutura

```
GET <index>/_search
{
  "aggs": {
    “<nomeAgregação>": {
      "<TipoAgregação>": {}
    }
  }
}
```


Semantix

Agregações Tipos

- Bucket:
  - Combinam os documentos resultantes em buckets
    - Buckets são criados
- Metric
  - Cálculos matemáticos feitos nos campos de documentos
    - São calculados em buckets
- Matrix
  - Operam em diversos campos produzindo uma matriz de resultado (matrix_stats)
- Pipeline
  - Agrega a saída de outras agregações

Agregações Tipos

- Buckets
  - Conjunto de documento formado por critérios
    - Data
    - Intervalo
    - Atributo
  - Ex.
    - Range
    - Date_range
    - Ip_ranges
    - Geo_distance
    - Significant_terms
    - Etc
- Métricas
  - Operações matemáticas
    - Um valor de saída
    - Ex.
      - Avg
      - Sum
      - Min
      - Max
      - Cardinality
      - Value_count
      - Etc
  - Operações matemáticas
    - N valores de saída
    - Ex.
      - Stats
      - Percentiles
      - Percentile_ranks
      - Etc

Agregações de Métricas

Semantix

Agregações Exemplo - Avg

- Média do campo qtd

```
GET cliente/_search
{
  "query": { ... },
  "aggs": {
    "media": {
      "avg": {
        "field": "qtd"
      }
    }
  }
}
```

Agregações Exemplo – Sum com limitação de escopo

- Visualizar apenas o resultado da agregação, ou uma parte dos resultados
  - size

```
GET cliente/_search
{
  "query": { ... },
  "size": 0,
  "aggs": {
    “soma": {
      “sum": {
        "field": "qtd"
      }
    }
  }
}
```

Agregações Exemplo - Stats

- Várias estatísticas com apenas uma requisição

```
GET cliente/_search
{
  "query": { ... },
  "aggs": {
    “estatistica": {
      "stats": {
        "field": "qtd"
      }
    }
  }
}
```

- Estatísticas
  - "count"
  - "min"
  - "max"
  - "avg"
  - "sum"

Agregações Exemplo – Min e Max

- Valor mínimo e máximo do campo qtd

```
GET cliente/_search
{
  "aggs": {
    "minimo": {
      "min": { "field": "qtd" }
    },
    "maximo": {
      "max": {  "field": "qtd" }
    }
  }
}
```

Agregações Exemplo – Cardinalidade

- Contar valores únicos
  - O resultado pode não ser preciso para grandes datasets
    - HyperLogLog++ algorithm
    - Precisão x Velocidade

```
GET cliente/_search
{
  "size“:0,
  "aggs": {
    “quantidade_cidades": {
      “cartinality": {
        "field": “cidade.keyword“
      }
    }
  }
}
```


Semantix

Agregações Exemplo – Mediana

- Mediana do campo qtd

```
GET cliente/_search
{
  "query": { ... },
  "aggs": {
    "mediana": {
      “median": {
        "field": "qtd"
      }
    }
  }
}
```

ERRO – Não existe está operação

Agregações de Buckets

Semantix

Agregações Exemplo – Separar em porcentagem

- Mediana do campo qtd

```
GET cliente/_search
{
  "aggs": {
    "mediana": {
      “percentiles": {
        "field": "qtd"
      }
    }
  }
}
```

- Median is 445.0

- Resposta do Elasticsearch

```
{
  ...
  "aggregations": {
    "load_time_outlier": {
      "values" : {
        "1.0": 5.0,
        "5.0": 25.0,
        "25.0": 165.0,
        "50.0": 445.0,
        "75.0": 725.0,
        "95.0": 945.0,
        "99.0": 985.0
}}}}
```

Agregações Exemplo – Separar em porcentagem

- Mediana do campo qtd

```
GET cliente/_search
{
  "aggs": {
    "media": {
      “percentiles": {
        "field": "qtd“,
        "percents": [25, 50, 75, 100]
      }
    }
  }
}
```

- Median is 445.0

Agregações Exemplo – Tempo

- Agrupar valores por um intervalo
  - date_histogram

```
GET logs_servico/_search {
 "size": 0,
 "aggs": {
  "logs_por_dia": {
   "date_histogram": {
    "field": "@timestamp",
    "calendar_interval ": "day“
   }
} } }
```

- Opções:
  - “calendar_interval”: “month”
  - “fixed_interval”:”10m”
    - Medidas:
      - ms, s, m, h, d, w, M, q, y

Agregações Exemplo – Tempo

- Agrupar valores por um valor específico
  - histogram

```
GET logs_servico/_search {
 "size": 0,
 "aggs": {
  "logs_cada_100ms: {
   "histogram": {
    "field": “runtime_ms",
    "interval": 100
   }
  }
 }
}
```


Semantix

Agregações Exemplo – Intervalo

```
GET cliente/_search
{
  "query": { ... },
  "aggs": {
    "intervalo": {
      "range": {
        "field": "qtd",
        "ranges": [
          { "to": 5},
          { "from": 5, "to": 20 },
          { "from": 20 }
        ]
      }
    }
  } }
```

Agregações Exemplo – Intervalo de Data

```
GET cliente/_search
{
  "query": { ... },
  "aggs": {
    "intervalo_data": {
      "date_range": {
        "field": "data",
        "ranges": [
          { "from": 2019-01-01, "to": 2019-05-01 }
        ]
      }
    }
  }
}
```

Agregações Exemplo – Atributo

- Especificar o campo e a quantidade de valores
  - Valores com a maior relevancia
- Ex. As 5 maiores cidades que visitaram o site

```
GET logs_servico/_search {
 "size": 0,
 "aggs": {
  “cidades_views": {
   "terms": {
    "field": “cidade.keyword",
    "size": 5
   }
  }
 }
}
```

Exercícios Agregações

- Realizar os exercícios no índice bolsa

1. Calcular a média do campo volume
2. Calcular a estatística do campo close
3. Visualizar os documentos do dia 2019-01-01 à 2019-03-01. (hits = 9)
4. Visualizar os documentos do dia 2019-04-01 até agora. (hits = 3)
5. Calcular a estatística do campo open do período do dia 2019-04-01 até agora
6. Calcular a mediana do campo open
7. Contar a quantidade de documentos agrupados por ano
8. Contar a quantidade de documentos de 2 anos atrás até hoje

Semantix

Obrigado!

Alguma pergunta?

Você pode me encontrar em:

rodrigo.augusto@semantix.com.br

GET SMARTER