Eléonor KIOULOU
Paul RUNAVOT
Adèle MONTOYA

TP3 ELASTICSEARCH

INSTALLATION ELASTICSEARCH :

Pull Elasticsearch (for mac) :

```
S8/NoSql/Elasticsearch via  desktop-linux via  v3.9.13 took 9s
❯ docker-compose up
```

docker-compose 〉No Selection

```yaml
version: '3'
services:
  elasticsearch:
    image: docker.elastic.co/elasticsearch/elasticsearch:7.17.0
    environment:
      - discovery.type=single-node
    ports:
      - 9200:9200
    volumes:
      - ./elasticsearch-data:/usr/share/elasticsearch/data
  kibana:
    image: docker.elastic.co/kibana/kibana:7.17.0
    ports:
      - 5601:5601
    environment:
      - ELASTICSEARCH_URL=http://elasticsearch:9200

```

Modifying the json to match the elasticsearch format :

```
S8/NoSql/Elasticsearch via  desktop-linux via  v3.9.13 took 9s
❯ python3 app.py
```

```python
import json
from datetime import datetime

input_file = "companies2.json"
output_file = "output.json"

def convert_to_iso8601_fromstr(date_str):
    # Convert the date string to a datetime object
    date_obj = datetime.strptime(date_str, '%a %b %d %H:%M:%S %Z %Y')
    # Convert the datetime object to a string in ISO 8601 format
    return date_obj.isoformat()

def convert_to_iso8601_fromobj(date_obj):
    # Extract the date and time string from the input string
    date_time_str = date_obj["$date"].split(".")[0] + "Z"
    # Convert the date and time string to a datetime object
    date_obj = datetime.strptime(date_time_str, '%Y-%m-%dT%H:%M:%SZ')
    return date_obj.isoformat()

with open(input_file, "r") as f_input, open(output_file, "w") as f_output:
    for i, line in enumerate(f_input):
        # Parse the JSON object
        obj = json.loads(line)
        # Convert the date format
        # It may be an object or a string
        if isinstance(obj["created_at"], str):
            obj["created_at"] = convert_to_iso8601_fromstr(obj["created_at"])
        else:
            obj["created_at"] = convert_to_iso8601_fromobj(obj["created_at"])
        if isinstance(obj["updated_at"], str):
            obj["updated_at"] = convert_to_iso8601_fromstr(obj["updated_at"])
        else:
            obj["updated_at"] = convert_to_iso8601_fromobj(obj["updated_at"])
        # Remove the id from the document
        _id = obj.pop("_id", None)
        # Write the index line
        f_output.write('{"index":{"_index":"companies","_id":%d}}\n' % i)
        # Write the data line
        f_output.write(json.dumps(obj) + "\n")
```

=>

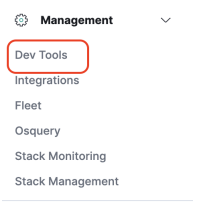
```
{"index":{"_index":"companies","_id":0}}
{"name": "Vidyo", "permalink": "vidyo", "crunchbase_url": "http://www.crunchbase.com/company/vidyo", "
{"index":{"_index":"companies","_id":1}}
{"name": "WeGame", "permalink": "wegame", "crunchbase_url": "http://www.crunchbase.com/company/wegame"
{"index":{"_index":"companies","_id":2}}
{"name": "Fraud Sciences", "permalink": "fraud-sciences", "crunchbase_url": "http://www.crunchbase.com
{"index":{"_index":"companies","_id":3}}
{"name": "Intronis", "permalink": "intronis", "crunchbase_url": "http://www.crunchbase.com/company/int
{"index":{"_index":"companies","_id":4}}
{"name": "IRSeeK", "permalink": "irseek", "crunchbase_url": "http://www.crunchbase.com/company/irseek"
{"index":{"_index":"companies","_id":5}}
{"name": "CrossLoop", "permalink": "crossloop", "crunchbase_url": "http://www.crunchbase.com/company/c
{"index":{"_index":"companies","_id":6}}
...
```

Import JSON :

```
S8/NoSql/Elasticsearch via  desktop-linux via  v3.9.13 took 9s
❯ curl -XPUT localhost:9200/_bulk -H "Content-Type: application/json" --data-binary @output.json
```

http://localhost:9200/

http://localhost:5601/



Eléonor KIOULOU
Paul RUNAVOT
Adèle MONTOYA

SIMPLE QUERY :

Get the companies that have more than 1000 employees:

```
GET /companies/_search
{
  "_source": ["name", "number_of_employees"],
  "query": {
    "range": {
      "number_of_employees": {
        "gt": 1000
      }
    }
  }
}
```


```
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "74",
  "_score" : 1.0,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 80000,
    "name" : "Apple"
  }
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "118",
  "_score" : 1.0,
  "_source" : {
    "number_of_employees" : 8000,
    "name" : "NetApp"
  }
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "220",
  "_score" : 1.0,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 51000,
    "name" : "Motorola Solutions"
  }
```

```
  "_score" : 1.0,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 4400,
    "name" : "Expedia"
  }
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "361",
  "_score" : 1.0,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 15500,
    "name" : "Experian"
  }
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "449",
  "_score" : 1.0,
  "_ignored" : [
    "overview.keyword",
    "video_embeds.embed_code.keyword"
  ],
  "_source" : {
    "number_of_employees" : 2000,
    "name" : "Netflix"
```


Etc…

Retrieve all companies that were founded before 2000, sorted in descending order of number of employees:

```
GET /companies/_search
{
  "_source": ["name", "founded_year", "number_of_employees"],
  "sort": {"number_of_employees": "desc"},
  "query": {
    "range": {
      "founded_year": {
        "lt": 2000
      }
    }
  }
}
```


```
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "5776",
  "_score" : null,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 405000,
    "founded_year" : 1847,
    "name" : "Siemens"
  },
  "sort" : [
    405000
  ]
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "8105",
  "_score" : null,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 320000,
    "founded_year" : 1933,
    "name" : "Toyota"
  },
  "sort" : [
    320000
  ]
},
```

```
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "1255",
  "_score" : null,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 300000,
    "founded_year" : 1998,
    "name" : "PayPal"
  },
  "sort" : [
    300000
  ]
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "13172",
  "_score" : null,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "number_of_employees" : 227000,
    "founded_year" : 1985,
    "name" : "Nippon Telegraph and Telephone Corporation"
  },
  "sort" : [
    227000
  ]
},
```


Etc...

Eléonor KIOULOU
Paul RUNAVOT
Adèle MONTOYA

Retrieve information from companies whose name contains "Google" and which were founded after 2005:

```
GET /companies/_search
{
  "_source": ["name", "founded_year"],
  "query": {
    "bool": {
      "must": [
        {
          "match": {
            "name": "Google"
          }
        },
        {
          "range": {
            "founded_year": {
              "gte": 2005
            }
          }
        }
      ]
    }
  }
}
```


```
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "5072",
      "_score" : 6.5603886,
      "_source" : {
        "founded_year" : 2008,
        "name" : "Google And Blog"
      }
    },
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "6637",
      "_score" : 6.5603886,
      "_ignored" : [
        "overview.keyword"
      ],
      "_source" : {
        "founded_year" : 2005,
        "name" : "Google Earth Blog"
      }
    },
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "7543",
      "_score" : 6.5603886,
      "_ignored" : [
        "overview.keyword"
      ],
      "_source" : {
        "founded_year" : 2005,
        "name" : "Google Earth Blog"
      }
```



```
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "11552",
      "_score" : 6.5603886,
      "_source" : {
        "founded_year" : 2008,
        "name" : "SEO Google Guru"
      }
    },
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "10952",
      "_score" : 5.6597548,
      "_source" : {
        "founded_year" : 2008,
        "name" : "Google Friend Connect Directory"
      }
    }
  ]
}
}
```


Récupérer les entreprises ayant une adresse e-mail avec "@gmail.com" dans leur champ email_address:

```
GET /companies/_search
{
  "_source": ["name", "email_address"],
  "query": {
    "wildcard": {
      "email_address": "*gmail.com"
    }
  }
}
```


```
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "113",
      "_score" : 1.0,
      "_ignored" : [
        "overview.keyword"
      ],
      "_source" : {
        "email_address" : "whatsopen.com@gmail.com",
        "name" : "WhatsOpen"
      }
    },
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "153",
      "_score" : 1.0,
      "_ignored" : [
        "overview.keyword"
      ],
      "_source" : {
        "email_address" : "dongthaicuatoi@gmail.com",
        "name" : "Widgetbox"
      }
    },
```



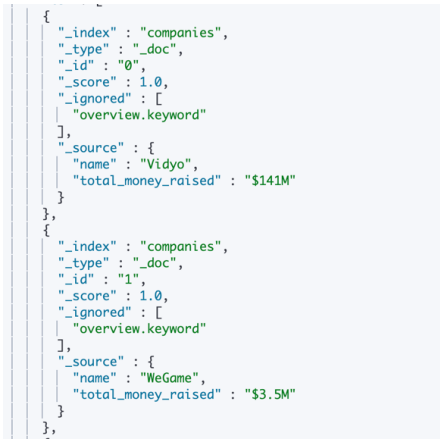
```
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "189",
      "_score" : 1.0,
      "_ignored" : [
        "overview.keyword"
      ],
      "_source" : {
        "email_address" : "cagster75@gmail.com",
        "name" : "Pornotube"
      }
    },
    {
      "_index" : "companies",
      "_type" : "_doc",
      "_id" : "232",
      "_score" : 1.0,
      "_ignored" : [
        "overview.keyword"
      ],
      "_source" : {
        "email_address" : "youporn@gmail.com",
        "name" : "YouPorn"
      }
    },
```


Etc

Eléonor KIOULOU
Paul RUNAVOT
Adèle MONTOYA

Retrieve companies that have received total funding over $10 million:

```
GET /companies/_search
{
  "_source": ["name", "total_money_raised"],
  "query": {
    "range": {
      "total_money_raised": {
        "gt": "10M"
      }
    }
  }
}
```



Etc…

Retrieve companies that have been acquired:

```
GET /companies/_search
{
  "_source": ["name"],
  "query": {
    "exists": {
      "field": "acquisition"
    }
  }
}
```

Etc..

Eléonor KIOULOU
Paul RUNAVOT
Adèle MONTOYA

COMPLEX QUERY:

Search for all companies that raised more than $10 million in funding and were founded after 2010:

```
GET /companies/_search
{
  "_source": ["name", "founded_year","total_money_raised"],
  "query": {
    "bool": {
      "must": [
        { "range": { "founded_year": { "gte": 2010 } } },
        { "range": { "total_money_raised": { "gte": 10000000 } } }
      ]
    }
  }
}
```


```
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "97",
  "_score" : 2.0,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "founded_year" : 2012,
    "name" : "Pinger",
    "total_money_raised" : "$18.5M"
  }
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "153",
  "_score" : 2.0,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "founded_year" : 2012,
    "name" : "Widgetbox",
    "total_money_raised" : "$14.5M"
  }
},
```

```
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "157",
  "_score" : 2.0,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "founded_year" : 2011,
    "name" : "RazorGator",
    "total_money_raised" : "$58.8M"
  }
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "405",
  "_score" : 2.0,
  "_ignored" : [
    "overview.keyword",
    "video_embeds.embed_code.keyword"
  ],
  "_source" : {
    "founded_year" : 2013,
    "name" : "Advaliant",
    "total_money_raised" : "$100k"
  }
```


Search for all companies having "big data" in their description or in their keyword list, and having been founded between 2005 and 2015, sorted alphabetically by company name:

```
GET /companies/_search
{
  "_source": ["name", "founded_year", "description", "tag_list"],
  "query": {
    "bool": {
      "must": [
        { "range": { "founded_year": { "gte": 2005, "lte": 2015 } } },
        { "bool": {
            "should": [
              { "match": { "description": "big data" } },
              { "match": { "tag_list": "big data" } }
            ]
          }
        }
      ]
    }
  },
  "sort": [
    { "name.keyword": { "order": "asc" } }
  ]
}
```

Eléonor KIOULOU
Paul RUNAVOT
Adèle MONTOYA


```
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "10529",
  "_score" : null,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "tag_list" : "backup, data-backup, remote-backup, offsite
      tapeless, ohio, columbus, techcolumbus",
    "founded_year" : 2007,
    "name" : "3X Systems",
    "description" : "Business data backup appliances"
  },
  "sort" : [
    "3X Systems"
  ]
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "14379",
  "_score" : null,
  "_ignored" : [
    "overview.keyword",
    "video_embeds.embed_code.keyword"
  ],
  "_source" : {
    "tag_list" : "web-crawling, web-crawler, data",
    "founded_year" : 2008,
    "name" : "80legs",
    "description" : "Web Crawling Service"
  },
```



```
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "11690",
  "_score" : null,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "tag_list" : "affiliate-tools, affiliate-network-optimiza
      -money-online, affiliate-data-feeds, monetize-on-affili
    "founded_year" : 2008,
    "name" : "AffiliatesDrive",
    "description" : "Meta Affiliate Feed Server"
  },
  "sort" : [
    "AffiliatesDrive"
  ]
},
{
  "_index" : "companies",
  "_type" : "_doc",
  "_id" : "8367",
  "_score" : null,
  "_ignored" : [
    "overview.keyword"
  ],
  "_source" : {
    "tag_list" : "data, geolocation, database",
    "founded_year" : 2006,
    "name" : "AggData",
    "description" : "Location Data Extraction"
  },
  "sort" : [
    "AggData"
```


Etc…

HARD QUERY:

Search the most common category of companies with more than 100k employees.

```
GET /companies/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "range": {
            "number_of_employees": { "gte": 100000 }
          }
        }
      ]
    }
  },
  "aggs": {
    "top_categories": {
      "significant_terms": {
        "field": "category_code.keyword"
      }
    }
  }
}
```


```
{
  "took" : 86,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 17,
      "relation" : "eq"
    },
    "max_score" : 1.0,
```

Eléonor KIOULOU
Paul RUNAVOT
Adèle MONTOYA


```
"aggregations" : {
  "top_categories" : {
    "doc_count" : 17,
    "bg_count" : 18674,
    "buckets" : [
      {
        "key" : "hardware",
        "doc_count" : 6,
        "score" : 6.072951117398536,
        "bg_count" : 362
      },
      {
        "key" : "other",
        "doc_count" : 3,
        "score" : 0.41512987218704134,
        "bg_count" : 983
      }
    ]
  }
}
```