TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

ELASTICSEARCH

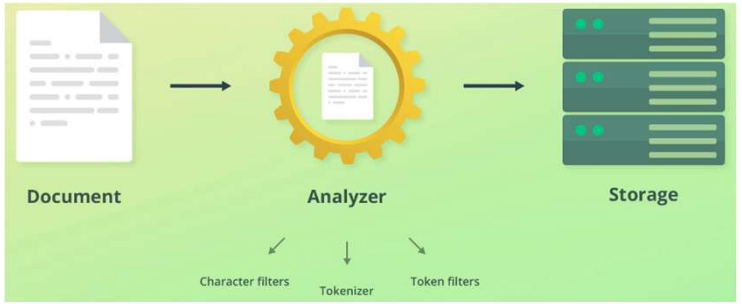
Lukáš Matějů

29.4.2024 | DPB

ČÁST I.:
OPAKOVÁNÍ

OPAKOVÁNÍ

- analýza
  - často také jako textová analýza
    - reálná aplikace jen na textová pole / hodnoty
  - vstupní data dokumentu v poli _source
    - ta ale nejsou přímo používána pro prohledávání
      - např. dlouhý popis produktu nelze efektivně prohledávat bez předzpracování
  - ➢ při indexaci jsou textová pole analyzována
    - o analýzu se stará analyzér skládající se ze tří komponent
      - znakový filtr (character filter)
      - tokenizér (tokenizer)
      - token filtr (token filter)
    - v základu standard analyzér
      - žádný znakový filtr
      - tokenizér (Unicode segmentation)
      - lowercase token filtr
    - ➢ výsledky analýzy uloženy v polích efektivních pro prohledávání



https://www.udemy.com/course/elasticsearch-complete-guide/

OPAKOVÁNÍ

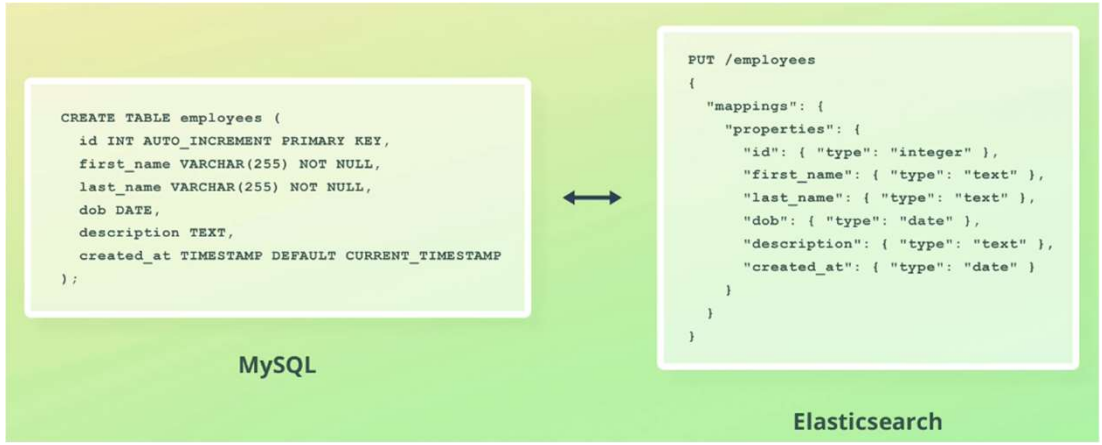
- invertovaný index
  - mapování mezi výrazy a dokumenty, které je obsahují
  - výrazy jsou řazeny abecedně
  - dokumenty jsou odkazovány pomocí _id

TERM	DOCUMENT #1	DOCUMENT #2	DOCUMENT #3
2	X	X	X
a	X	X	
around			X
bar	X	X	
but	X		
ducks	X		X
guys	X	X	
into	X	X	
lake			X
the	X		X
third	X		
walk	X		X
went		X	

https://www.udemy.com/course/elasticsearch-complete-guide/

OPAKOVÁNÍ

- mapování
  - definuje strukturu dokumentů, a jak jsou indexovány a ukládány
    - pole a jejich datové typy
  - zjednodušeně se dá přirovnat k relaci v relačních databázích
  - explicitní a dynamické
  - povoluje chybějící pole dokumentů



https://www.udemy.com/course/elasticsearch-complete-guide/

OPAKOVÁNÍ

- prohledávání – dotazovací typy
  - dotazy na úrovni termínů (term-level queries)

```
GET /products/_search
{
  "query": {
    "term": {
      "name": "lobster"
    }
  }
}
```

```
"hits" : {
  "total" : {
    "value" : 5,
    "relation" : "eq"
  },
  "max_score" : 6.035804,
  "hits" : [
    {
      "_index" : "products",
      "_type" : "_doc",
      "_id" : "19",
      "_score" : 6.035804,
      "_source" : {
        "name" : "Lobster - Live",
```

```
GET /products/_search
{
  "query": {
    "term": {
      "name": "Lobster"
    }
  }
}
```

```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 0,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  }
}
```

  - fulltextové dotazy (full-text queries)

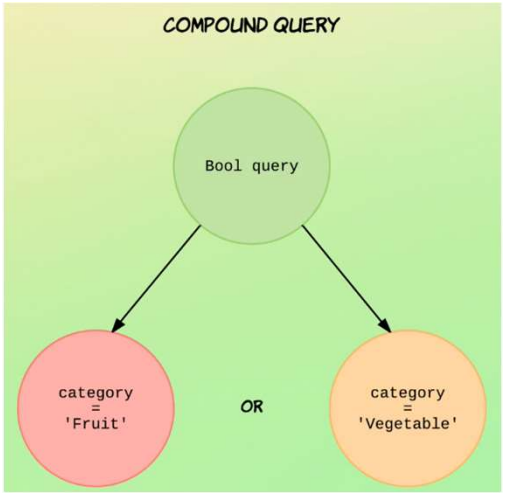
```
GET /products/_search
{
  "query": {
    "match": {
      "name": "Lobster"
    }
  }
}
```

```
"hits" : {
  "total" : {
    "value" : 5,
    "relation" : "eq"
  },
  "max_score" : 6.035804,
  "hits" : [
    {
      "_index" : "products",
      "_type" : "_doc",
      "_id" : "19",
      "_score" : 6.035804,
      "_source" : {
        "name" : "Lobster - Live",
```

OPAKOVÁNÍ

- prohledávání – dotazy s Booleovskou logikou
  - leaf queries
    - koncové dotazy
    - nejjednodušší dotazy provádějící jen jednu operaci
    - vyhledávají hodnotu v daném poli
    - např. term nebo match query
  - compound queries
    - tvoří složitější dotazy zaobalením leaf nebo dalších compound queries
    - např. dvě leaf query spojené přes bool query
    - ➢ využívají booleovskou logiku

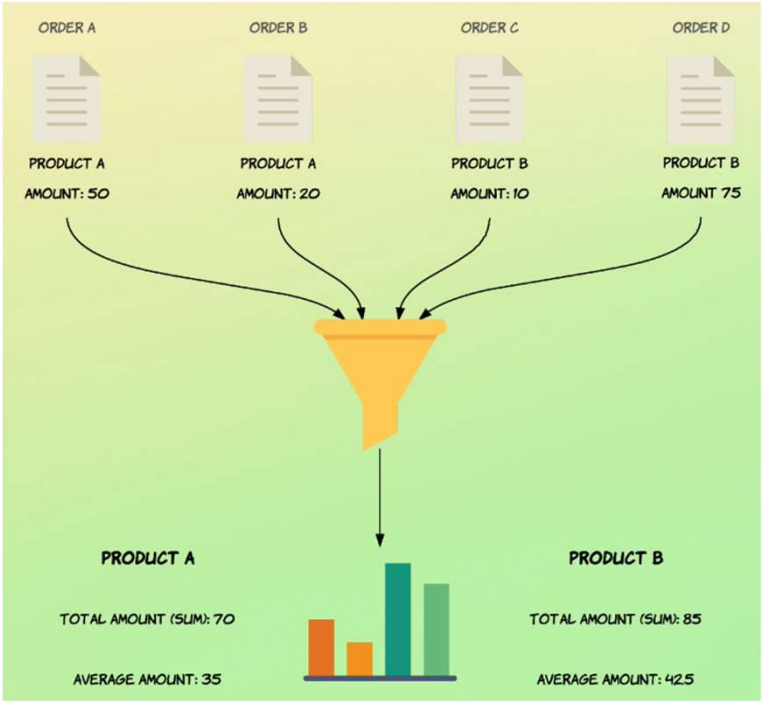
https://www.udemy.com/course/elasticsearch-complete-guide/



ČÁST II.:
ELASTICSEARCH
AGREGACE

AGREGACE

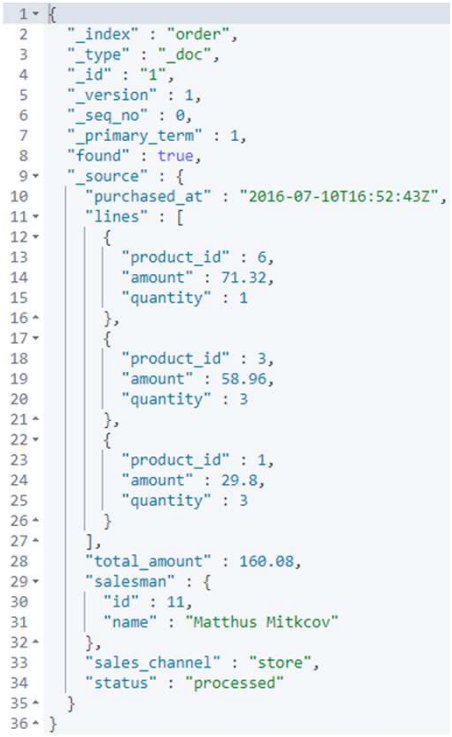
- způsob jak seskupit data a jak z nich získat statistiky a závěry



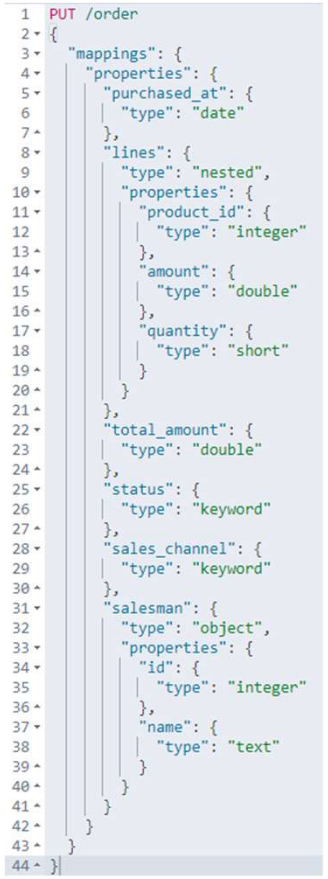
https://www.udemy.com/course/elasticsearch-complete-guide/

AGREGACE

- mapování a data pro následující ukázky
  - lines
    - specifikuje jednotlivé objednávky
    - nested objekt
    - id, amount, quantity
  - total_amount
    - celkové množství
  - status
    - keyword
    - status objednávky
  - sales_channel
    - kde bylo zboží prodáno
  - salesman
    - prodejce
    - id, jméno



```
{
  "_index" : "order",
  "_type" : "_doc",
  "_id" : "1",
  "_version" : 1,
  "_seq_no" : 0,
  "_primary_term" : 1,
  "found" : true,
  "_source" : {
    "purchased_at" : "2016-07-10T16:52:43Z",
    "lines" : [
      {
        "product_id" : 6,
        "amount" : 71.32,
        "quantity" : 1
      },
      {
        "product_id" : 3,
        "amount" : 58.96,
        "quantity" : 3
      },
      {
        "product_id" : 1,
        "amount" : 29.8,
        "quantity" : 3
      }
    ],
    "total_amount" : 160.08,
    "salesman" : {
      "id" : 11,
      "name" : "Matthus Mitkcov"
    },
    "sales_channel" : "store",
    "status" : "processed"
  }
}
```



```
PUT /order
{
  "mappings": {
    "properties": {
      "purchased_at": {
        "type": "date"
      },
      "lines": {
        "type": "nested",
        "properties": {
          "product_id": {
            "type": "integer"
          },
          "amount": {
            "type": "double"
          },
          "quantity": {
            "type": "short"
          }
        }
      },
      "total_amount": {
        "type": "double"
      },
      "status": {
        "type": "keyword"
      },
      "sales_channel": {
        "type": "keyword"
      },
      "salesman": {
        "type": "object",
        "properties": {
          "id": {
            "type": "integer"
          },
          "name": {
            "type": "text"
          }
        }
      }
    }
  }
}
```

AGREGACE

- agregace metrik (metric aggregation)
  - nejzákladnější agregace
  - dvě kategorie
    - single-value numeric metric aggregation
      - výstupem je jediná hodnota (např. průměr nebo suma)
    - multi-value numeric metric aggregation
      - výstupem je více hodnot
  - pole aggs
    - zajímá nás statistika všech prodejů přes všechny objednávky
      - celkový objem zboží na objednávku
      - minimum, maximum, průměr a suma celkového objemu

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "total_sales": {
      "sum": {
        "field": "total_amount"
      }
    },
    "avg_sale": {
      "avg": {
        "field": "total_amount"
      }
    },
    "min_sale": {
      "min": {
        "field": "total_amount"
      }
    },
    "max_sale": {
      "max": {
        "field": "total_amount"
      }
    }
  }
}
```

```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1000,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "max_sale" : {
      "value" : 281.77
    },
    "avg_sale" : {
      "value" : 109.20961
    },
    "min_sale" : {
      "value" : 10.27
    },
    "total_sales" : {
      "value" : 109209.61
    }
  }
}
```

AGREGACE

- agregace metrik (metric aggregation)
  - cardinality
    - počet unikátních hodnot pro dané pole
  - value_count
    - počet hodnot pro dané pole
  - stats
    - vybrané základní agregace v jednom balíčku

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "total_salesmen": {
      "cardinality": {
        "field": "salesman.id"
      }
    }
  }
}
```

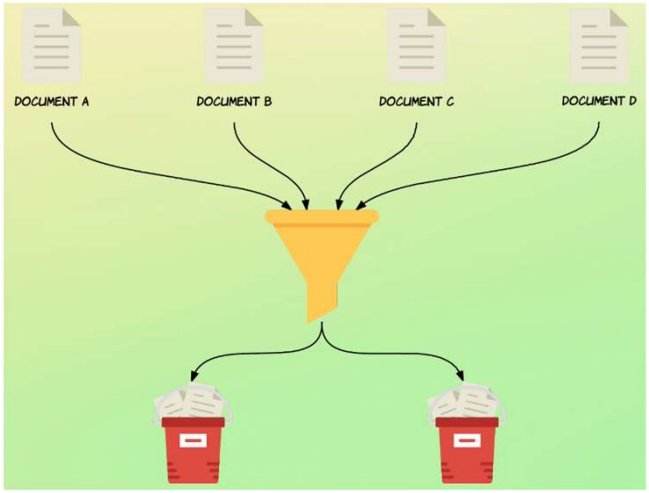
```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "values_count": {
      "value_count": {
        "field": "total_amount"
      }
    }
  }
}
```

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "amount_stats": {
      "stats": {
        "field": "total_amount"
      }
    }
  }
}
```

```
  "aggregations" : {
    "amount_stats" : {
      "count" : 1000,
      "min" : 10.27,
      "max" : 281.77,
      "avg" : 109.20961,
      "sum" : 109209.61
    }
  }
}
```

AGREGACE

- bucket agregace
  - vytváří skupiny (buckets) dokumentů
  - každý bucket má kritéria, která rozhodují, jestli do nich dokument spadá
  - agregace výrazů (terms)
    - vytváří bucket pro každý unikátní výraz
    - agregace podle statusu objednávky



https://www.udemy.com/course/elasticsearch-complete-guide/

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "status_terms": {
      "terms": {
        "field": "status",
        "missing": "N/A",
        "order": {
          "_key": "asc"
        }
      }
    }
  }
}
```

```
"aggregations" : {
  "status_terms" : {
    "doc_count_error_upper_bound" : 0,
    "sum_other_doc_count" : 0,
    "buckets" : [
      {
        "key" : "cancelled",
        "doc_count" : 196
      },
      {
        "key" : "completed",
        "doc_count" : 204
      },
      {
        "key" : "confirmed",
        "doc_count" : 192
      },
      {
        "key" : "pending",
        "doc_count" : 199
      },
      {
        "key" : "processed",
        "doc_count" : 209
      }
    ]
  }
}
```

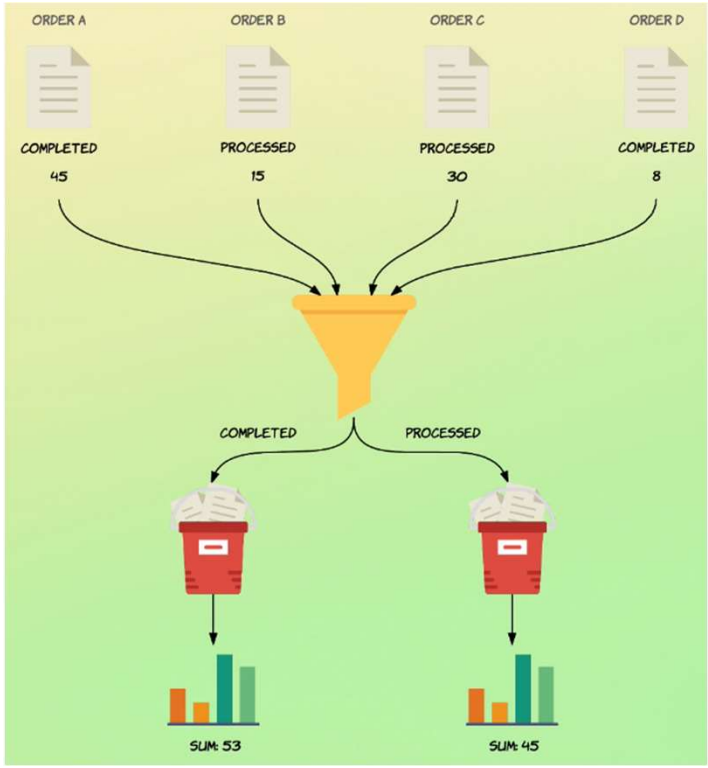
AGREGACE

- bucket agregace
  - vnořené (nested) agregace
    - také jako subagregace
    - vnoření dalšího aggs pole
    - ➢ další agregace na získaných buckets
    - statistika pro každý status

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "status_terms": {
      "terms": {
        "field": "status"
      },
      "aggs": {
        "status_stats": {
          "stats": {
            "field": "total_amount"
          }
        }
      }
    }
  }
}
```

```
{
  "key" : "processed",
  "doc_count" : 209,
  "status_stats" : {
    "count" : 209,
    "min" : 10.27,
    "max" : 281.77,
    "avg" : 109.30703349282295,
    "sum" : 22845.17
  }
},
```

```
{
  "key" : "completed",
  "doc_count" : 204,
  "status_stats" : {
    "count" : 204,
    "min" : 10.93,
    "max" : 260.59,
    "avg" : 113.54058823529411,
    "sum" : 23162.28
  }
},
```



https://www.udemy.com/course/elasticsearch-complete-guide/

AGREGACE

- bucket agregace
  - filter agregace
    - filter
    - filtrování dokumentů

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "low_value": {
      "filter": {
        "range": {
          "total_amount": {
            "lt": 50
          }
        }
      },
      "aggs": {
        "avg_amount": {
          "avg": {
            "field": "total_amount"
          }
        }
      }
    }
  }
}
```

```
{
  "took" : 6,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1000,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "low_value" : {
      "doc_count" : 164,
      "avg_amount" : {
        "value" : 32.59371951219512
      }
    }
  }
}
```

AGREGACE

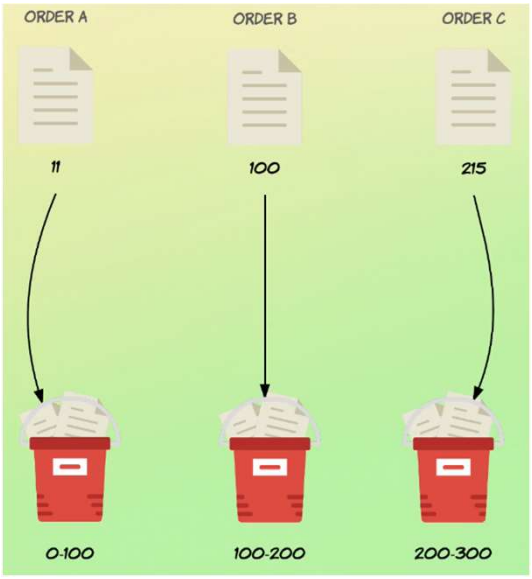
- bucket agregace
  - filters agregace
    - filters
    - definování pravidel určujících dokumenty patřící do vybraného bucket
    - např. bucket pro trička bude obsahovat jen trička

```
GET /recipes/_search
{
  "size": 0,
  "aggs": {
    "my_filter": {
      "filters": {
        "filters": {
          "pasta": {
            "match": {
              "title": "pasta"
            }
          },
          "spaghetti": {
            "match": {
              "title": "spaghetti"
            }
          }
        }
      },
      "aggs": {
        "avg_rating": {
          "avg": {
            "field": "ratings"
          }
        }
      }
    }
  }
}
```

```
{
  "took" : 4,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 21,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "my_filter" : {
      "buckets" : {
        "pasta" : {
          "doc_count" : 9,
          "avg_rating" : {
            "value" : 3.4125
          }
        },
        "spaghetti" : {
          "doc_count" : 4,
          "avg_rating" : {
            "value" : 2.3684210526315788
          }
        }
      }
    }
  }
}
```

AGREGACE

- bucket agregace
  - agregace rozsahů (range)
    - definice rozsahů, kde každý rozsah vytvoří bucket
    - dvě varianty
      - range – hodnoty
      - date_range – datumy



https://www.udemy.com/course/elasticsearch-complete-guide/

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "amount_distribution": {
      "range": {
        "field": "total_amount",
        "ranges": [
          {
            "to": 50
          },
          {
            "from": 50,
            "to": 100
          },
          {
            "from": 100
          }
        ]
      }
    }
  }
}
```

```
  "aggregations" : {
    "amount_distribution" : {
      "buckets" : [
        {
          "key" : "*-50.0",
          "to" : 50.0,
          "doc_count" : 164
        },
        {
          "key" : "50.0-100.0",
          "from" : 50.0,
          "to" : 100.0,
          "doc_count" : 347
        },
        {
          "key" : "100.0-*",
          "from" : 100.0,
          "doc_count" : 489
        }
      ]
    }
  }
}
```

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "purchased_ranges": {
      "date_range": {
        "field": "purchased_at",
        "ranges": [
          {
            "from": "2016-01-01",
            "to": "2016-01-01||+6M"
          },
          {
            "from": "2016-01-01||+6M",
            "to": "2016-01-01||+1y"
          }
        ]
      }
    }
  }
}
```

AGREGACE

- bucket agregace
  - agregace podle chybějících polí
    - missing
    - pole s hodnotou null nebo úplně chybící

```
POST /order/_doc/1001
{
  "total_amount": 100
}

POST /order/_doc/1002
{
  "total_amount": 200,
  "status": null
}
```

```
GET /order/_search
{
  "size": 0,
  "aggs": {
    "orders_without_status": {
      "missing": {
        "field": "status"
      }
    }
  }
}
```

```
{
  "took" : 1,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 1002,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "orders_without_status" : {
      "doc_count" : 2
    }
  }
}
```

AGREGACE

- bucket agregace
  - agregace vnořených objektů
    - nested

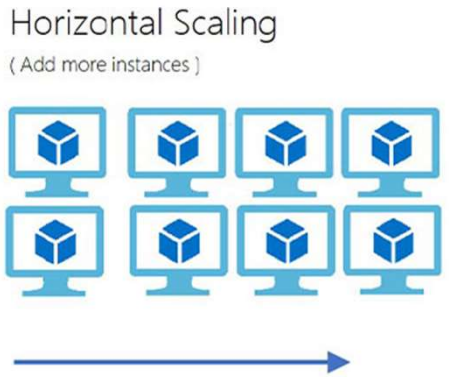
```
GET /departments/_search
{
  "size": 0,
  "aggs": {
    "employees": {
      "nested": {
        "path": "employees"
      },
      "aggs": {
        "minimum_age": {
          "min": {
            "field": "employees.age"
          }
        }
      }
    }
  }
}
```

```
{
  "took" : 3,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 2,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "employees" : {
      "doc_count" : 15,
      "minimum_age" : {
        "value" : 19.0
      }
    }
  }
}
```

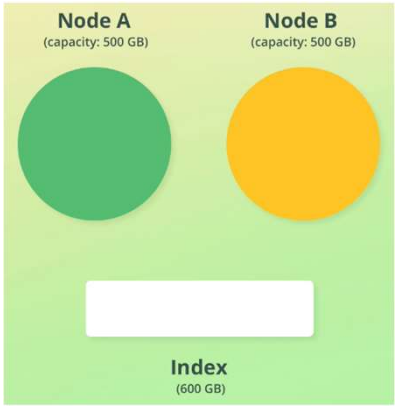
ČÁST III.:
ELASTICSEARCH
SHARDING A REPLIKACE

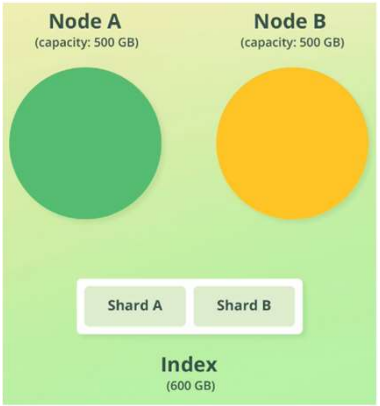
SHARDING

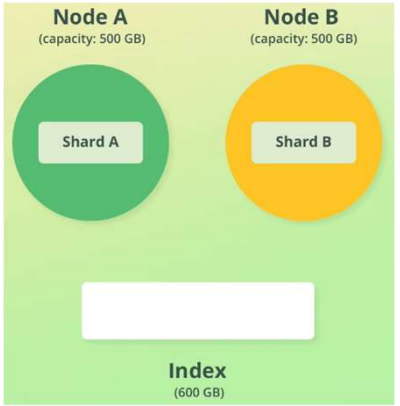
- rozdělení indexů na menší části
  - každá část je nazývána shard
  - ➢ prováděno na úrovni indexů
    - ne uzlů nebo clusterů
- ➢ horizontální škálování
  - škálování datového úložiště



https://www.webairy.com/horizontal-and-vertical-scaling/







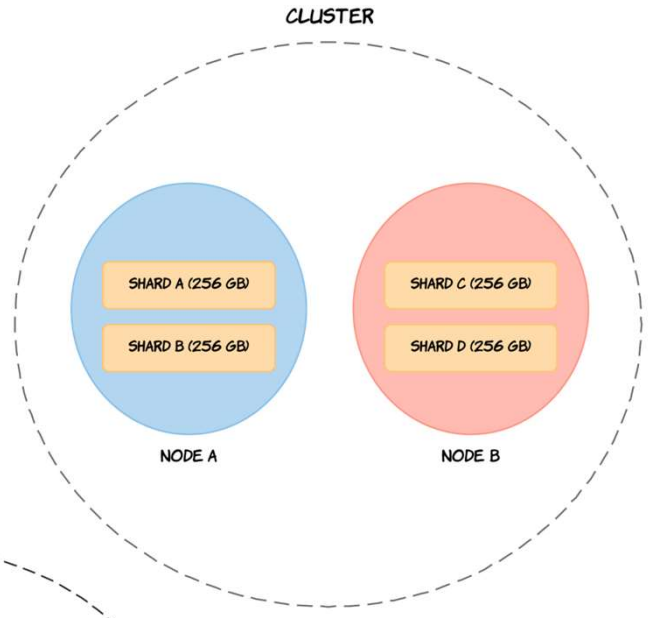
https://www.udemy.com/course/elasticsearch-complete-guide/

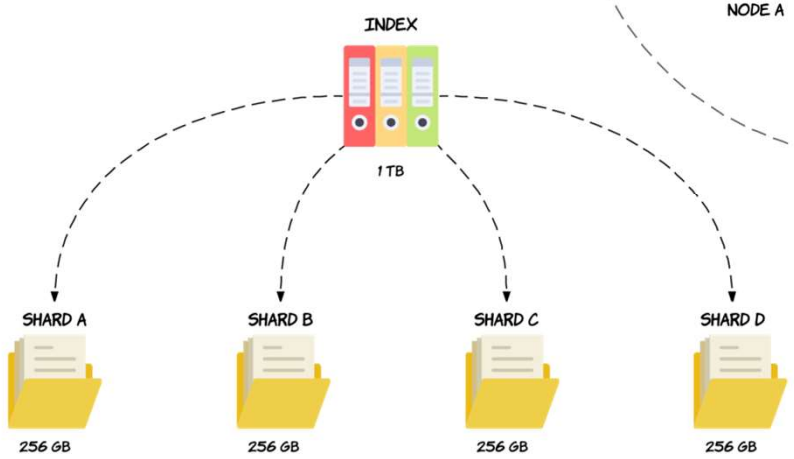
SHARDING

- každý shard
  - slouží víceméně jako samostatný index
    - Apache Lucene index
    - ➢ každý index je složen z alespoň jednoho Lucene indexu
  - nemá předdefinovanou velikost
    - roste s vkládanými dokumenty
  - může obsahovat až 2 miliardy dokumentů
- v základu každý index má jeden shard
  - možné konfigurovat
    - zvýšení počtu shardů pomocí Split API
    - snížení Shrink API
  - ➢ vhodné rozmyslet dopředu
    - neexistuje optimální počet shardů
    - závisí na konkrétní aplikaci

SHARDING

- proč shardovat?
  - ukládání více dokumentů
    - miliardy dokumentů v jednom indexu
  - rozdělení velkých indexů
    - rozdělení na části a uložení v uzlech
  - zlepšení výkonu
    - paralelní zpracování dotazů v shardech





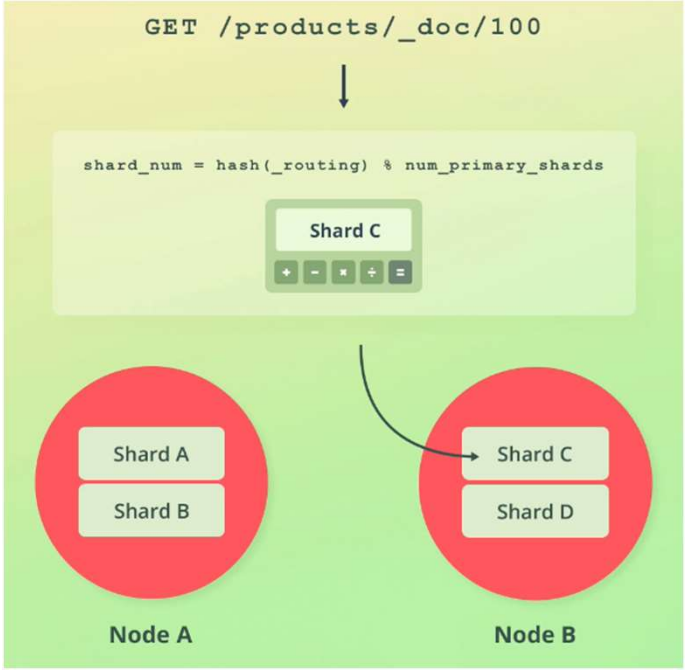
https://codingexplained.com/coding/elasticsearch/understanding-sharding-in-elasticsearch

ROUTING

- proces přiřazující dokumentu shard
  - určuje, kde bude dokument uložen
    - na jakém konkrétním shardu
  - udává, jak jsou dokumenty po zaindexování hledány
    - pro operace čtení, update nebo mazání
  - ➢ Elasticsearch využívá jednoduchý vzorec

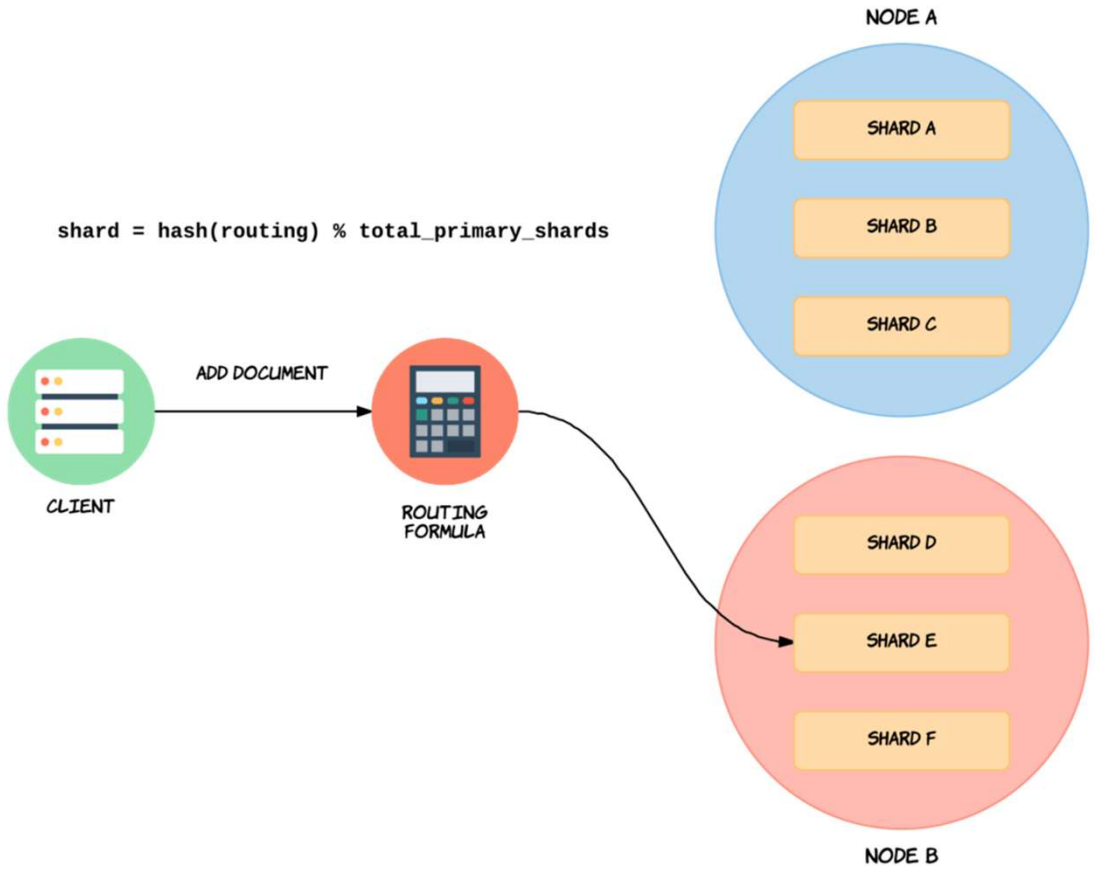
```
shard_num = hash(_routing) %
            num_primary_shards
```

    - _routing – odpovídá _id dokumentu
      - možnost vlastního nastavení
    - num_primary_shards – celkový počet shardů
  - zajišťuje rovnoměrné rozmístění dokumentů mezi shardy



https://www.udemy.com/course/elasticsearch-complete-guide/

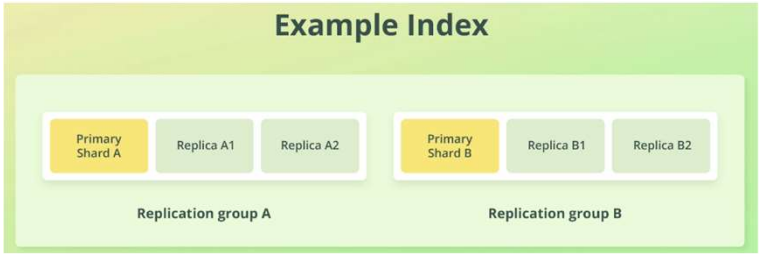
ROUTING



https://codingexplained.com/coding/elasticsearch/understanding-sharding-in-elasticsearch

REPLIKACE

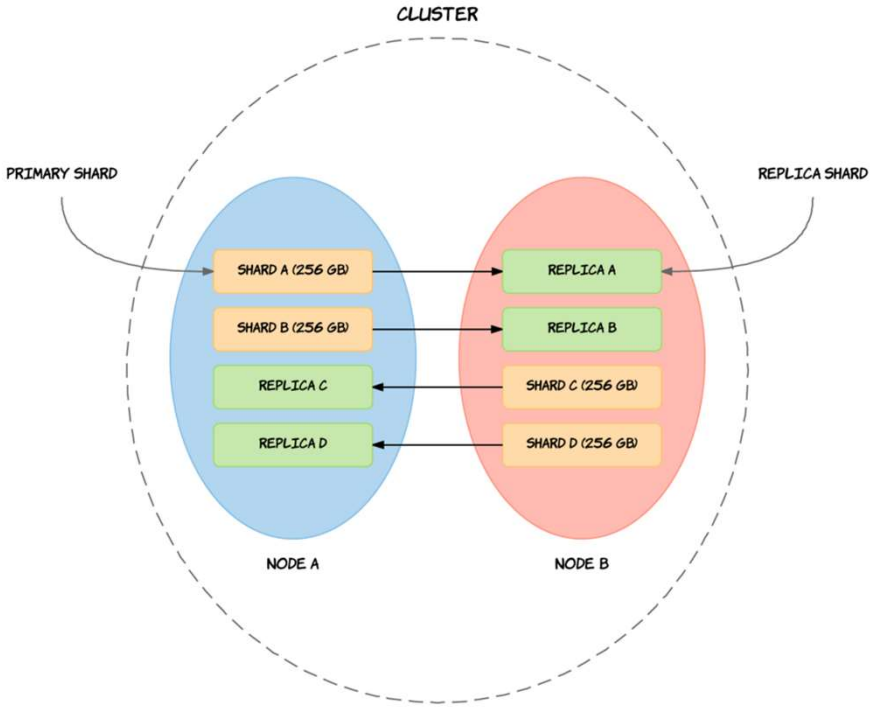
- automatická distribuce změn v originále do jeho kopií
- replikace v Elasticsearch
  - ochrana proti chybám (fault tolerance)
    - selhání hardware
  - nativní podpora, zapnuta automaticky
    - často obtížné nastavit v jiných databázích
  - definována na úrovni indexů
  - vytváří úplné kopie shardů
    - repliky shardů (replica shards)
    - původní shard se nazývá primární shard (primary shard)
    - primary shard a replica shards tvoří replication group
    - mohou plně obsluhovat dotazy
  - při vytváření indexu možnost nastavení počtu replik (1 v základu)



https://www.udemy.com/course/elasticsearch-complete-guide/

REPLIKACE

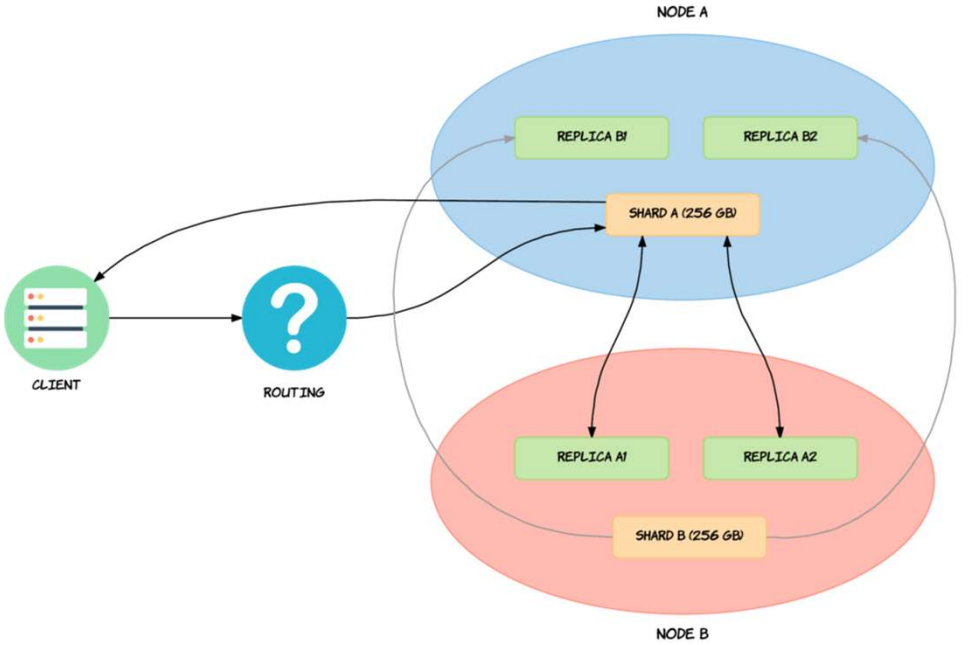
- replikace v Elasticsearch



https://codingexplained.com/coding/elasticsearch/understanding-replication-in-elasticsearch

REPLIKACE

- replikace v Elasticsearch
  - pro synchronizaci využíván model primary-backup
  - primární shard je vstupem pro indexovací operace
    - vkládání, update, mazání, …
  - primární shard operaci validuje a následně lokálně aplikuje na svá data
  - po dokončení je operace předána všem replikám
  - operace je paralelně provedena na všech replikách

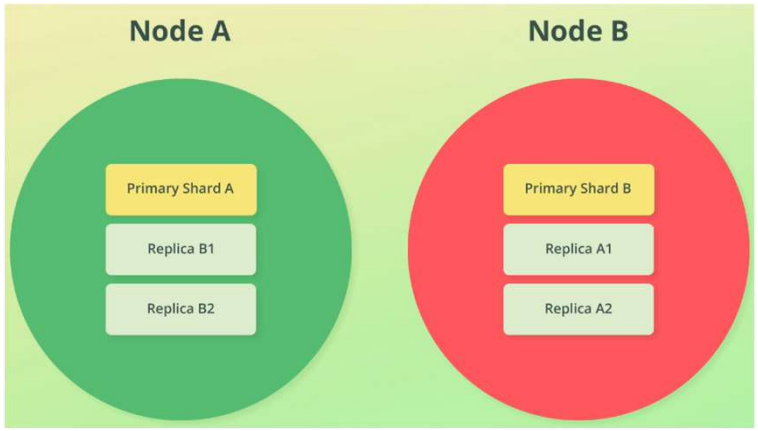


https://codingexplained.com/coding/elasticsearch/understanding-replication-in-elasticsearch

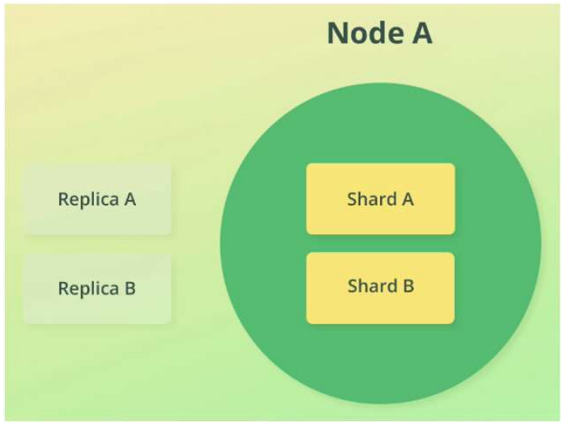
  - po potvrzení od všech replik informuje primární shard klienta

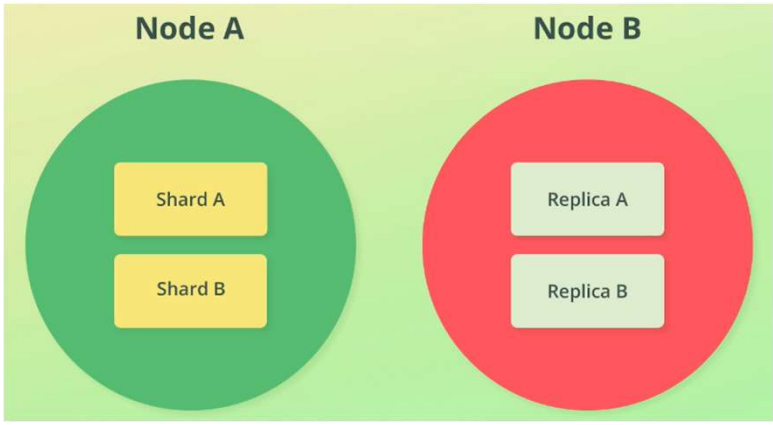
REPLIKACE

- replikace v Elasticsearch
  - jak zabránit ztrátě dat?
    - repliky nejsou na stejném uzlu jako primární shard
    - ➢ jsou potřeba alespoň dva uzly
    - ➢ Elasticsearch přidává repliky jen pro clustery s alespoň dvěma uzly
  - ➢ zvyšuje dostupnost indexu



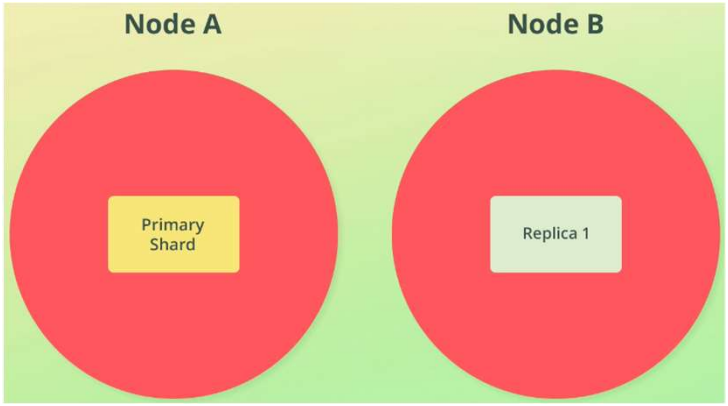
https://www.udemy.com/course/elasticsearch-complete-guide/



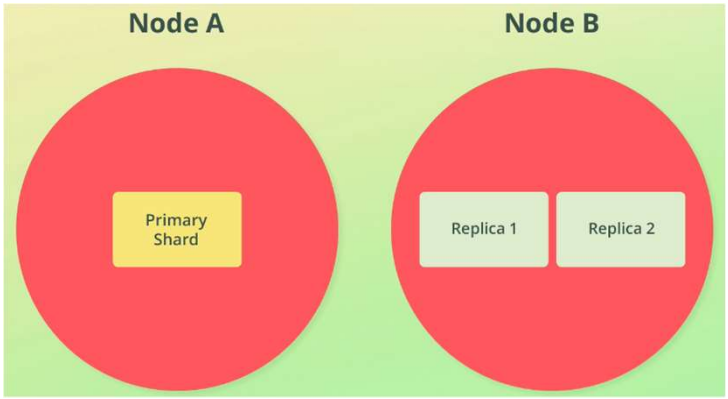


REPLIKACE

- replikace v Elasticsearch
  - ideální počet replik?
    - závisí na aplikaci
    - je možnost data obnovit z jiného zdroje?
    - je problémem krátká nedostupnost?
    - ➢ jedna replika v případě, že ztráta dat není katastrofická
    - ➢ alespoň dvě pro klíčové systémy
      - např. aplikace pro nemocnice, …
  - ➢ může také zvyšovat propustnost dotazování
    - repliky plně obsluhují dotazy
    - obsluhují je paralelně
    - ➢ dochází ke zvýšení zpracovaných dotazů ve stejný čas

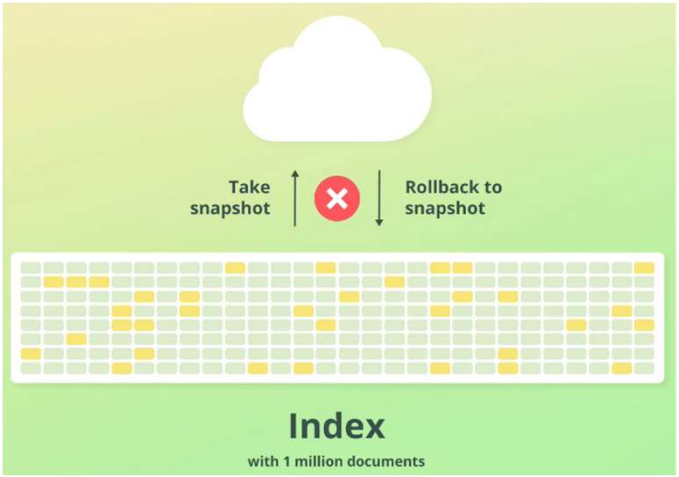


https://www.udemy.com/course/elasticsearch-complete-guide/



SNAPSHOTY

- snímek databáze určený pro zálohu stavu databáze
- snímky v Elasticsearch
  - umožňují obnovu databáze do stavu, ve kterém byl snímek pořízen
  - mohou být definované nad indexy ale i nad celým clusterem
  - export dat do souboru
- snapshoty vs. replikace
  - snímky slouží pro zálohu stavu
    - nezajišťují dostupnost aktuálních dat
  - replikace pro vysokou dostupnost
    - jen pro aktuální data
    - poskytuje obnovu při výpadku uzlu
    - neumožňuje obnovu historických dat



https://www.udemy.com/course/elasticsearch-complete-guide/

PŘIDÁNÍ UZLU (PRO VÝVOJ)

- kompletně změněno od verze 8.0
  - dříve jen další instance Elasticsearch
  - nyní založené na tzv. enrollment token
  - nejprve je potřeba enrollment token vygenerovat na libovolném uzlu

```
bin/elasticsearch-create-enrollment-token -s node
```

  - následně je enrollment token předán při spuštění uzlu

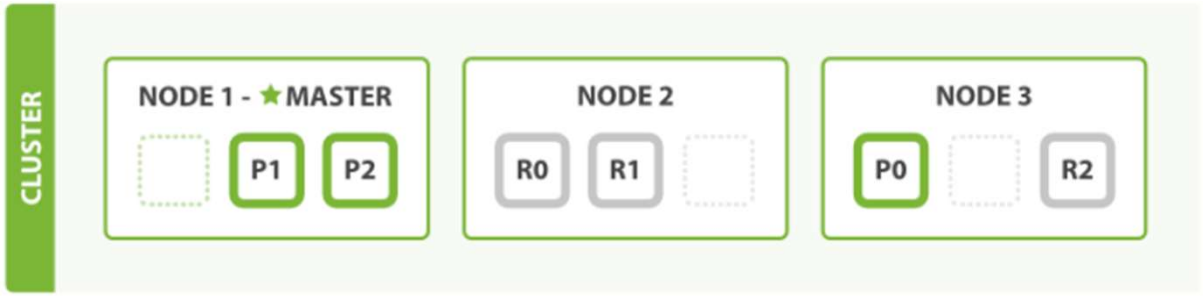
```
bin/elasticsearch --enrollment-token <enrollment-token>
```

    - certifikáty a klíče jsou automaticky vygenerovány do config/certs
  - enrollment token může být předán i dalším uzlům pro další rozšíření clusteru
  - jen uzly na stejném stroji se mohou připojit bez dalšího nastavení

https://www.elastic.co/guide/en/elasticsearch/reference/current/configuring-stack-security.html#stack-enroll-nodes

PŘIDÁNÍ UZLU (PRO VÝVOJ)

- pro praktické nasazení v clusteru potřeba i další nastavení
  - transport.host musí být nastaven pro připojení uzlů z různých zdrojů
- nelze použít pro Elastic Cloud
  - přidání/odebrání uzlů v interface Elastic Cloud
  - další uzly jsou samozřejmě placené
  - přidělení replica shards automaticky



https://www.elastic.co/guide/en/elasticsearch/reference/current/add-elasticsearch-nodes.html

PŘIDÁNÍ UZLU (PRO VÝVOJ)

```
[2020-11-18T14:10:41,619][INFO ][o.e.c.r.a.AllocationService] [node-1] updating number_of_replicas to [1] for indices [ilm-history-3-000001, .apm-custom-link, .apm-agent-configuration, .kibana_1, .kibana_task_manager_1, .kibana-event-log-7.10.0-000001]
[2020-11-18T14:10:41,621][INFO ][o.e.c.s.MasterService    ] [node-1] node-join[{node-2}{lmhdqVKZQuuMO6qsIELH8Q}{vSeL76bNTrC1I75EX4OumA}{127.0.0.1}{127.0.0.1:9301}{cdhilmrstw}{ml.machine_memory=34315870208, ml.max_open_jobs=20, xpack.installed=true, transform.node=true} join existing leader], term: 797, version: 153, delta: added {{node-2}{lmhdqVKZQuuMO6qsIELH8Q}{vSeL76bNTrC1I75EX4OumA}{127.0.0.1}{127.0.0.1:9301}{cdhilmrstw}{ml.machine_memory=34315870208, ml.max_open_jobs=20, xpack.installed=true, transform.node=true}}
[2020-11-18T14:10:42,449][INFO ][o.e.c.s.ClusterApplierService] [node-1] added {{node-2}{lmhdqVKZQuuMO6qsIELH8Q}{vSeL76bNTrC1I75EX4OumA}{127.0.0.1}{127.0.0.1:9301}{cdhilmrstw}{ml.machine_memory=34315870208, ml.max_open_jobs=20, xpack.installed=true, transform.node=true}}, term: 797, version: 153, reason: Publication{term=797, version=153}
[2020-11-18T14:10:57,570][INFO ][o.e.c.r.a.AllocationService] [node-1] Cluster health status changed from [YELLOW] to [GREEN] (reason: [shards started [[.kibana-event-log-7.10.0-000001][0]]]).
```

```
PUT /test

GET /_cluster/health

GET /_cat/shards?v
```

```
{
  "cluster_name" : "elasticsearch",
  "status" : "green",
  "timed_out" : false,
  "number_of_nodes" : 2,
  "number_of_data_nodes" : 2,
  "active_primary_shards" : 7,
  "active_shards" : 14,
  "relocating_shards" : 0,
  "initializing_shards" : 0,
  "unassigned_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" : 100.0
}
```

index	shard	prirep	state	docs	store	ip	node
.kibana_task_manager_1	0	r	STARTED	5	122.9kb	127.0.0.1	node-2
.kibana_task_manager_1	0	p	STARTED	5	132.2kb	127.0.0.1	node-1
ilm-history-3-000001	0	r	STARTED			127.0.0.1	node-2
ilm-history-3-000001	0	p	STARTED			127.0.0.1	node-1
.apm-agent-configuration	0	r	STARTED	0	208b	127.0.0.1	node-2
.apm-agent-configuration	0	p	STARTED	0	208b	127.0.0.1	node-1
.kibana-event-log-7.10.0-000001	0	r	STARTED	3	16.4kb	127.0.0.1	node-2
.kibana-event-log-7.10.0-000001	0	p	STARTED	3	16.4kb	127.0.0.1	node-1
test	0	r	STARTED	0	208b	127.0.0.1	node-2
test	0	p	STARTED	0	208b	127.0.0.1	node-1
.apm-custom-link	0	r	STARTED	0	208b	127.0.0.1	node-2
.apm-custom-link	0	p	STARTED	0	208b	127.0.0.1	node-1
.kibana_1	0	r	STARTED	26	10.4mb	127.0.0.1	node-2
.kibana_1	0	p	STARTED	26	10.4mb	127.0.0.1	node-1

ROLE UZLŮ

- možno definovat pro každý uzel až několik rolí
  - podle aplikace samotného uzlu
- uzly podle rolí
  - master
    - jeden, často dedikovaný
    - v případě více uzlů s touto rolí jen jeden je zvolen masterem
    - zodpovědnost za vytváření a mazání indexů, přiřazování shardů uzlům, …
  - data
    - ukládání dat
    - dotazy nad daty
  - ingest
    - ingest pipeline – série kroků prováděná při indexaci dokumentů
    - ➢ přidání dokumentů do indexů
      - ➢ má právo dokumenty měnit

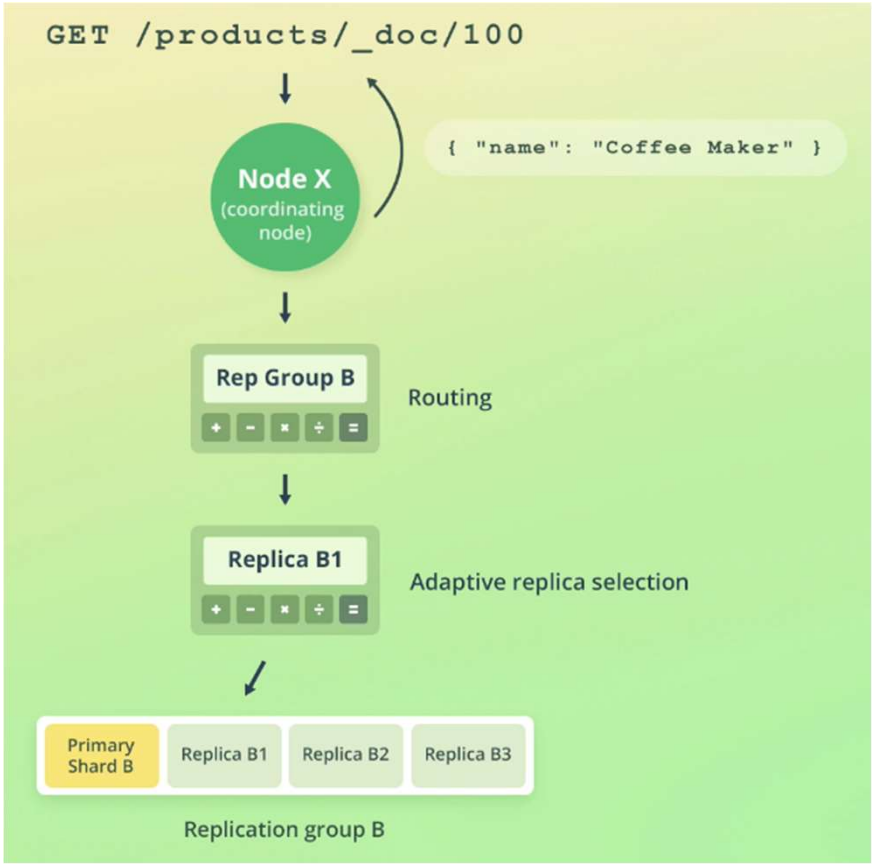
ROLE UZLŮ

- uzly podle rolí
  - machine learning
    - možnost provádět úlohy strojového učení
  - coordination
    - distribuce dotazů a agregace výsledků
  - voting-only
    - účast na volbě nového master uzlu
    - nemůže být master

```
GET /_cat/nodes?v
```

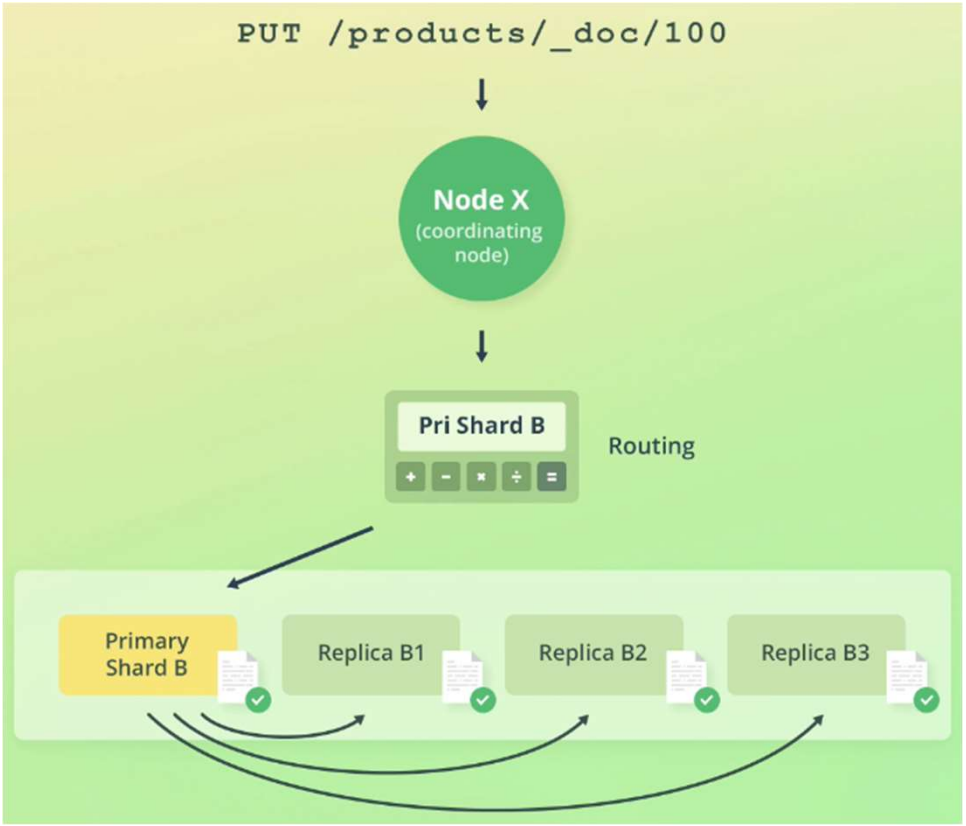
```
ip        heap.percent ram.percent cpu load_1m load_5m load_15m node.role  master name
127.0.0.1           21          43   8                          cdhilmrstw -      node-2
127.0.0.1           32          43   8                          cdhilmrstw *      node-1
```

ČTENÍ DOKUMENTU PODLE ID



https://www.udemy.com/course/elasticsearch-complete-guide/

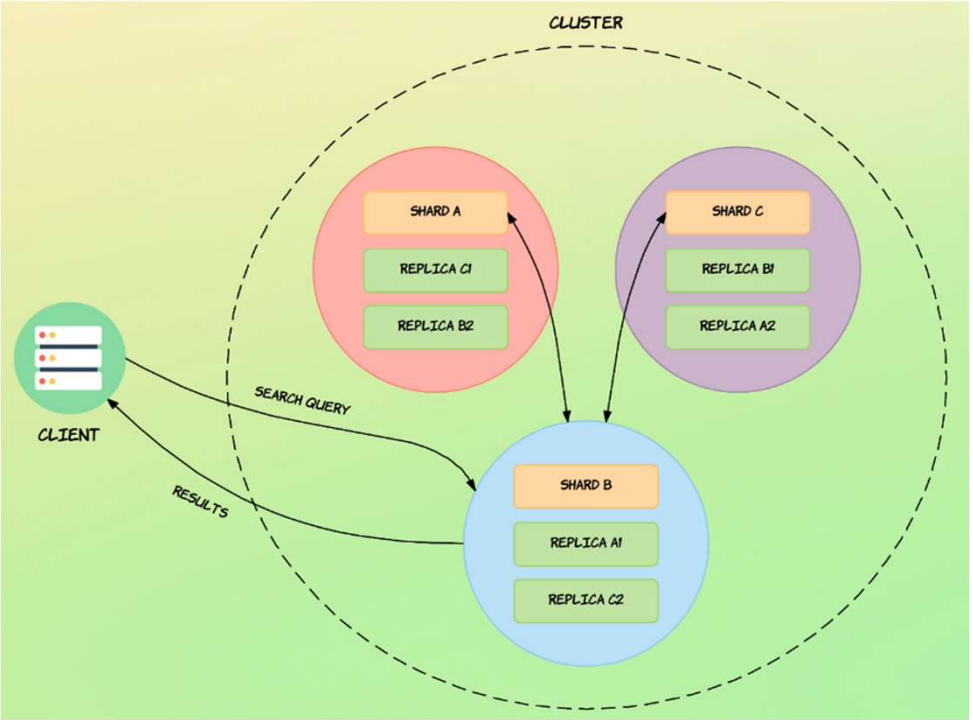
ZÁPIS DOKUMENTU PODLE ID



https://www.udemy.com/course/elasticsearch-complete-guide/

VYHLEDÁVÁNÍ

- princip
  - dotaz dorazí od klienta na jeden z uzlů (modrý)
  - ten se stává koordinačním uzlem
  - broadcastem přeposílá dotaz ostatním shardům
    - primárním i replikám
  - shardy odpovídají
  - koordinační uzel sloučí výsledky a vrací je klientovi
  - ➢ liší se výrazně od čtení podle _id

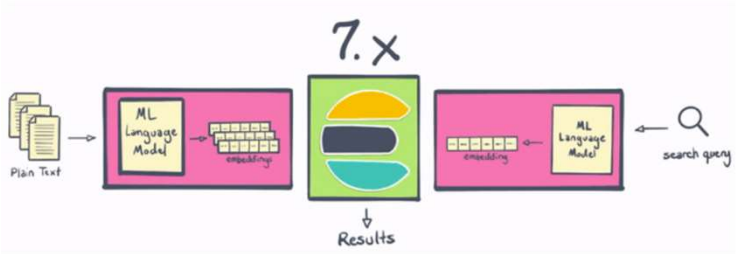


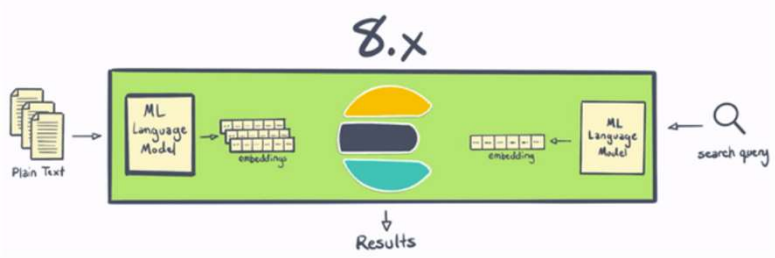
https://www.udemy.com/course/elasticsearch-complete-guide/

ČÁST IV.:
ELASTICSEARCH 8.0

CO JE NOVÉHO?

- nová verze 8.0 vydaná 02/2022 přináší různá vylepšení
  - vyšší rychlost a menší paměťové nároky
  - nativní podpora pro zpracování přirozeného jazyka (NLP)
    - analýza sentimentu, rozpoznávání entit, klasifikace textu, …
    - možnost přímého použití PyTorch modelů
  - vyšší relevance výsledků díky nativnímu vektorovému vyhledávání
  - vyšší rychlost a škálovatelnost vyhledávání
    - algoritmus approximate nearest neighbor search





https://www.elastic.co/blog/whats-new-elastic-8-0-0

CO JE NOVÉHO?

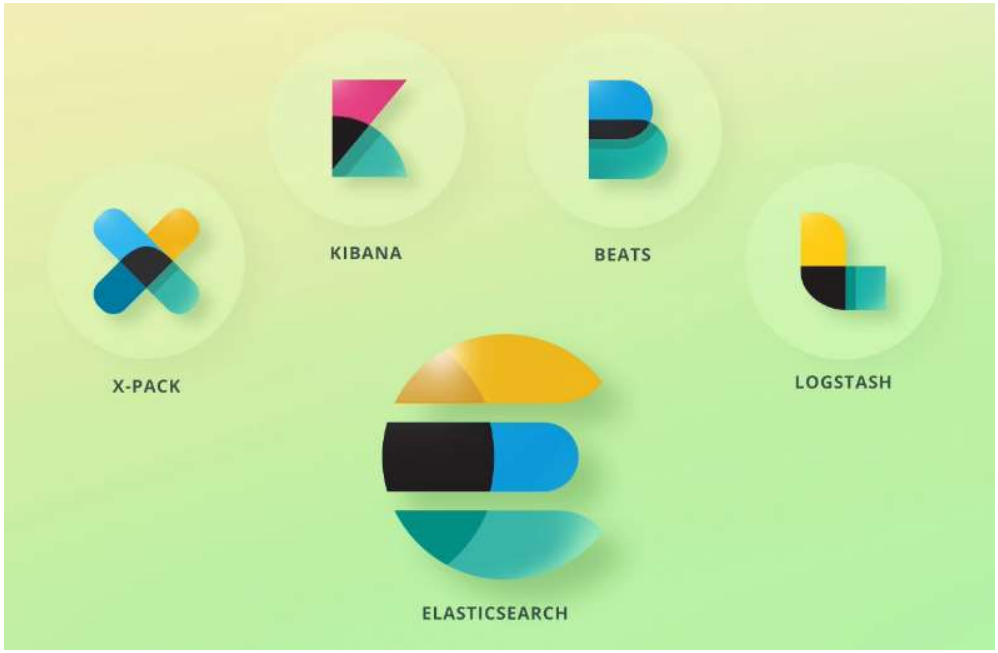
- nová verze 8.0 vydaná 02/2022 přináší různá vylepšení
  - vylepšená integrace AWS
  - vylepšená bezpečnost
    - autentizace, autorizace, šifrovaná komunikace (mezi uzly i mezi Elasticsearch a Kibana)
    - zapnuté a nastavené v základu

- informace ke zpětné kompatibilitě REST API
  - verze 8.0 přináší určité změny
    - ideální implementovat
  - ale nabízí také zpětnou kompatibilitu díky 7.x API compatibility headers
    - možné použít před přechodem na nové API
  - ➢ výrazné problémy tedy nejsou

ČÁST V.:
ELASTIC STACK

ELASTIC STACK

- technologie vyvinuté a spravované Elastic BV
  - silná synergie mezi technologiemi
    - často používané společně s Elasticsearch



https://www.udemy.com/course/elasticsearch-complete-guide/

KIBANA

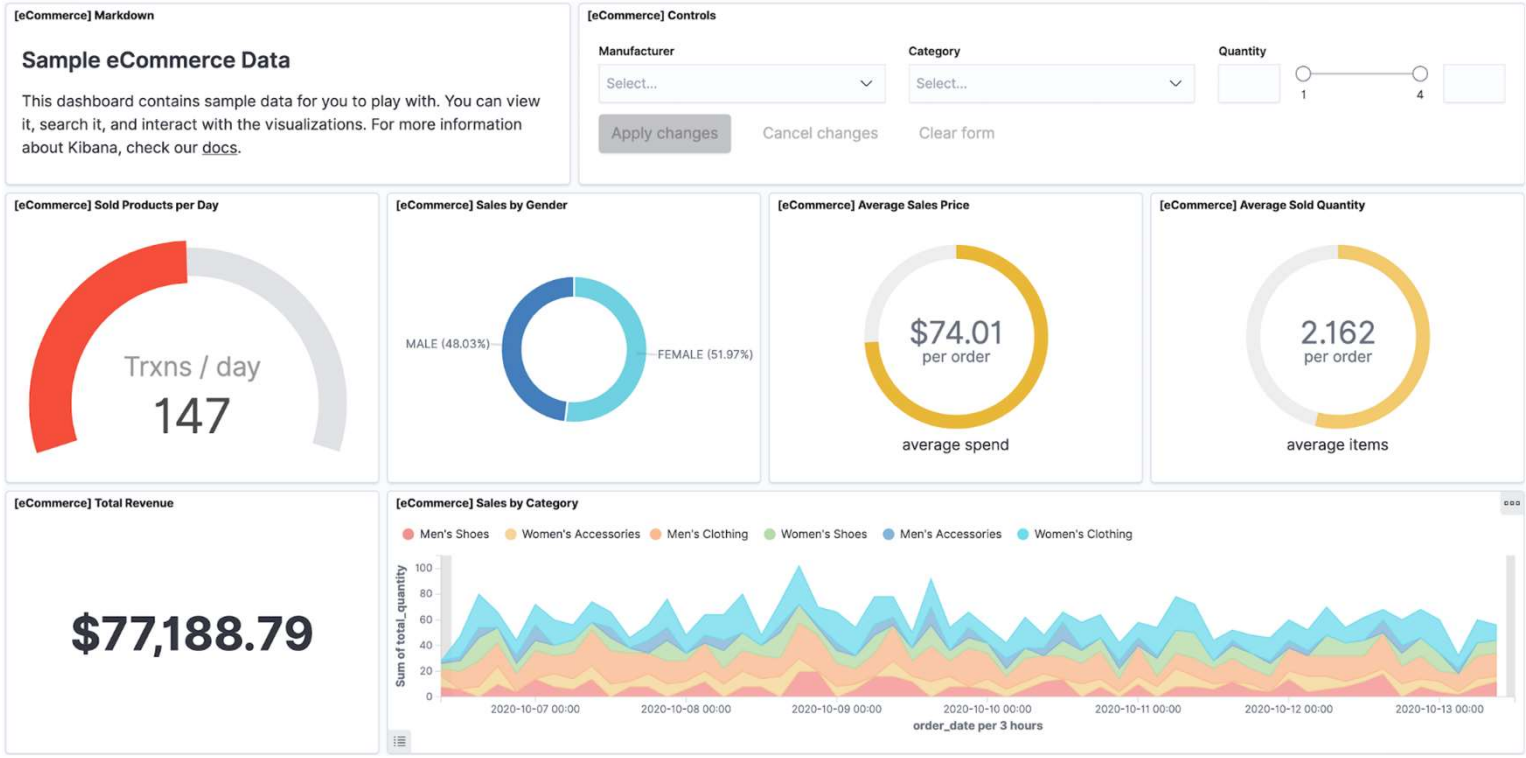
- platforma pro analýzu a vizualizaci dat v Elasticsearch
  - webový interface pro data uložená v Elasticsearch
  - vestavěná real-time vizualizace
    - grafy, mapy, …
    - dashboards
  - nastavení strojového učení
    - predikce, detekce anomálií
  - může spravovat určité části Elasticsearch a Logstash
    - autentifikace, autorizace

https://www.udemy.com/course/elasticsearch-complete-guide/

KIBANA

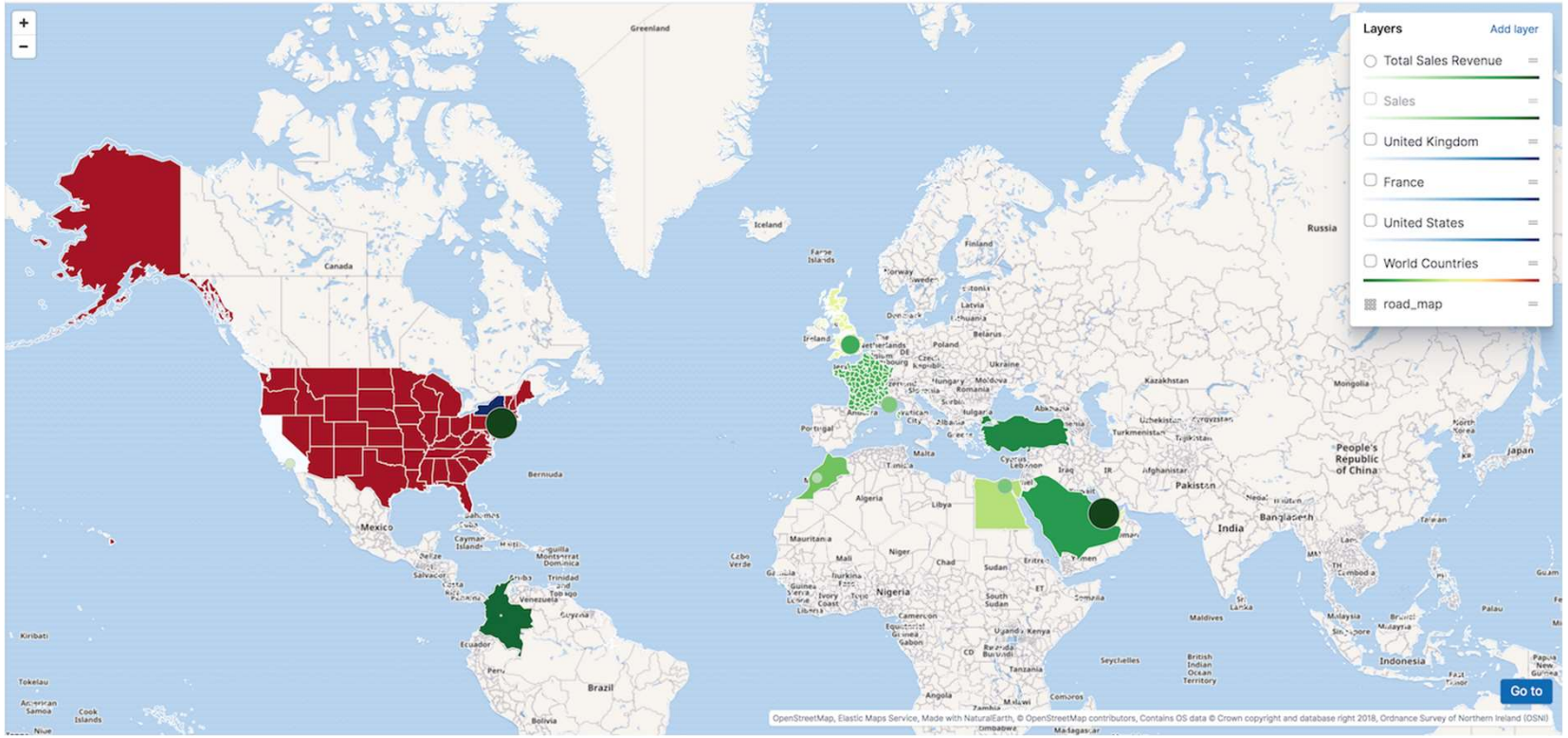
https://www.elastic.co/guide/en/kibana/current/kuery-query.html

KIBANA



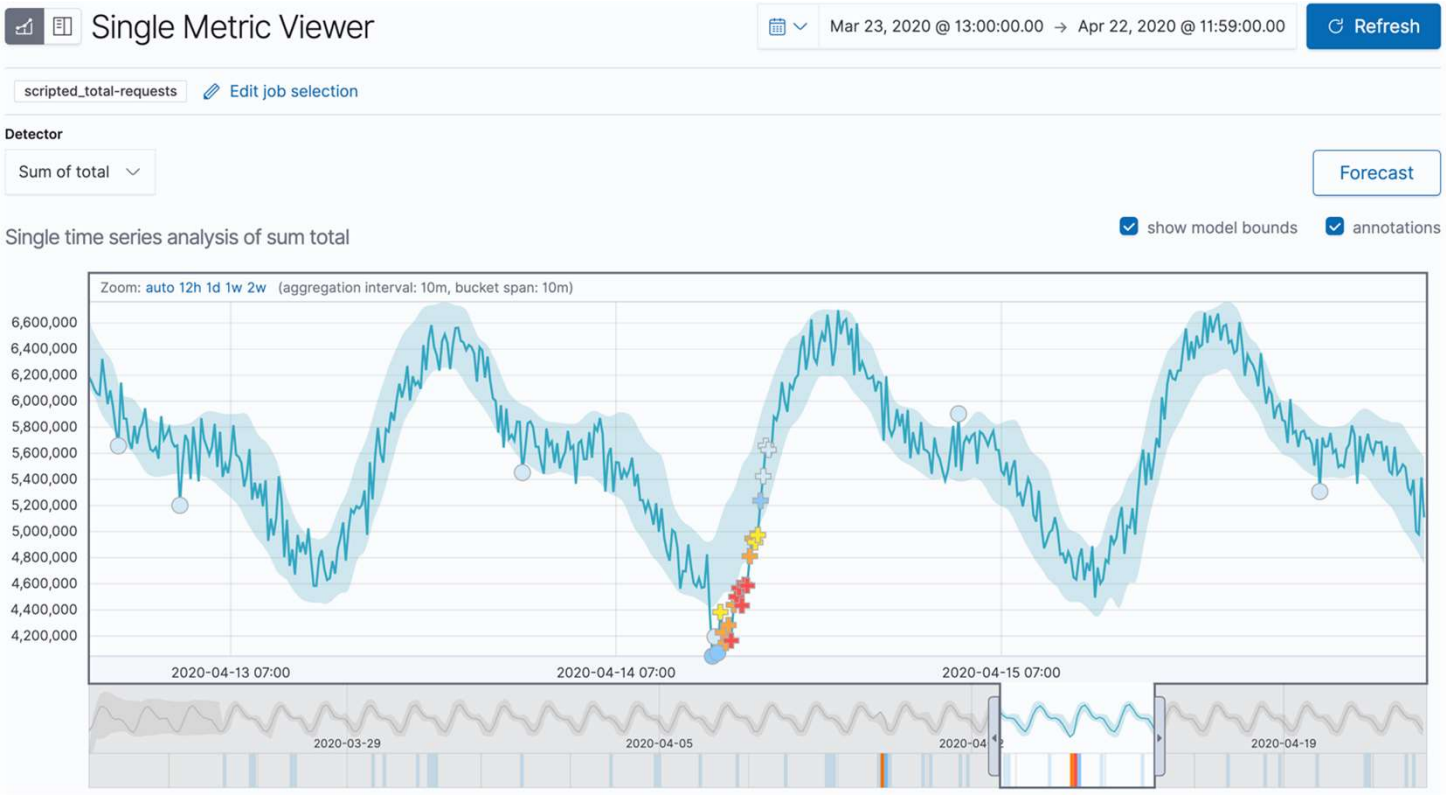
https://www.elastic.co/guide/en/kibana/current/dashboard.html

KIBANA



https://www.elastic.co/guide/en/kibana/current/maps.html

KIBANA

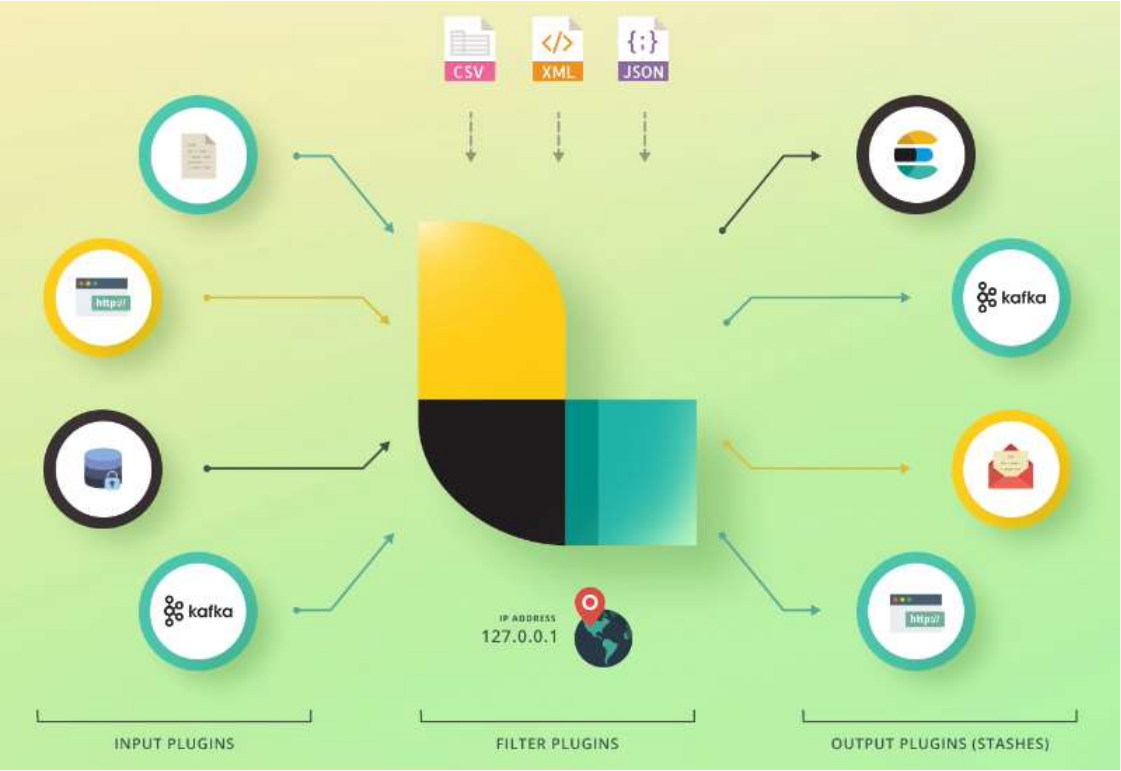


https://www.elastic.co/guide/en/kibana/current/xpack-ml-anomalies.html

LOGSTASH

- původně nástroj na zpracování logů a předávání do Elasticsearch
- dnes roura (pipeline) na zpracování dat
  - obecnější nástroj
  - hlavní princip
    - data vstupují do Logstash a jsou považována za obecné události
      - mohou být cokoliv (serverové logy, objednávky, zákazníci, zprávy z chatu, …)
    - Logstash události zpracovává a předává dál
      - může být kamkoliv (Elasticsearch, e-mail, web, …)
  - roura se skládá ze tří částí
    - vstupy, filtry, výstupy
    - každá část využívá pluginy pro práci s událostmi
      - vstupní pluginy – soubory, data přes HTTP, relační databáze, …
      - filtrační pluginy – parsování XML a JSON, obohacení dat (IP adresa, geo-informace), …
      - výstupní pluginy – stashes, místa, kam jsou události posílány dál
  - horizontální škálování

LOGSTASH



https://www.udemy.com/course/elasticsearch-complete-guide/

LOGSTASH

- roura na zpracování dat
  - definovaný vlastní značkovací formát (podobnost s JSON)
  - podporuje ale i podmíněné výrazy
    - dynamická roura

```
input {
    file {
        path => "/path/to/apache_access.log"
    }
}

filter {
    if [request] in ["/robots.txt", "/favicon.ico"] {
        drop { }
    }
}

output {
    file {
        path => "%{type}_%{+yyyy_MM_dd}.log"
    }
}
```

LOGSTASH

https://www.udemy.com/course/elasticsearch-complete-guide/

X-PACK

- balík funkcí do Elasticsearch a Kibana
  - zabezpečení
    - autentifikace a autorizace
    - integrace s poskytovateli autentifikací
    - nastavení uživatelských práv
  - monitoring
    - sledování výkonu Elastic Stack
    - využití procesorů, pamětí, disků, …
    - upozornění v případě něčeho neobvyklého
  - upozornění
  - reporty
    - export dat (a vizualizace) z Kibany do různých formátů (CSV, PDF, …)
    - generované na základě požadavků, podle plánu nebo po splnění určených podmínek
    - upravitelný vzhled

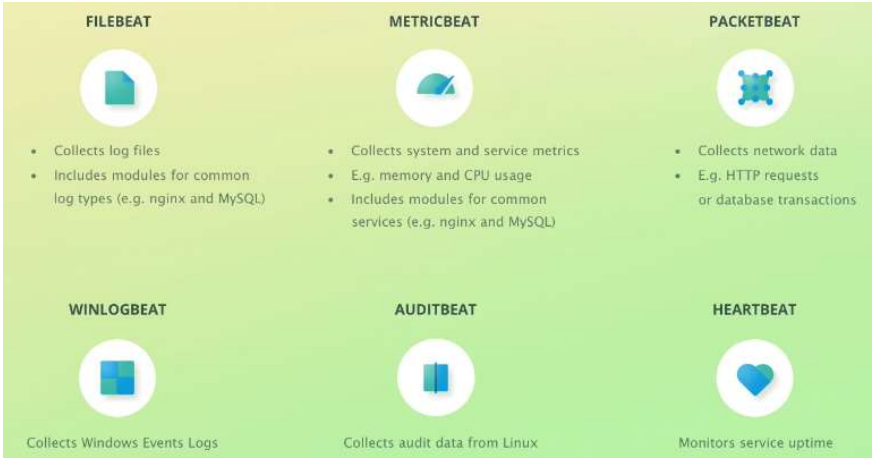
X-PACK

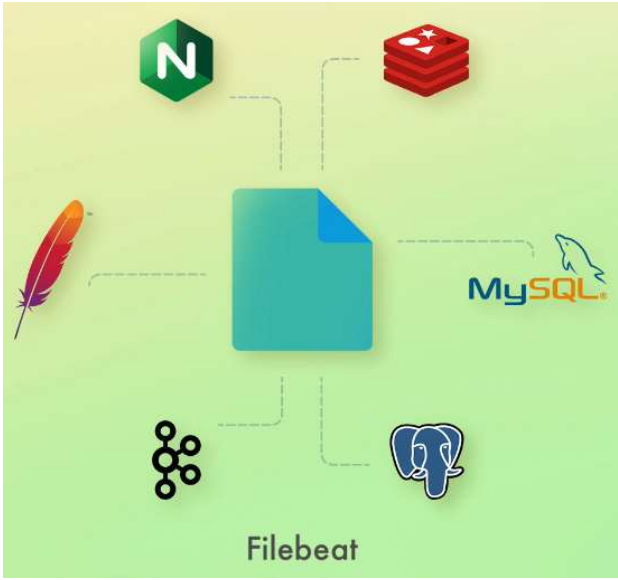
- balík funkcí do Elasticsearch a Kibana
  - strojové učení
    - funkcionalita dodána v X-Pack
    - interface v Kibaně
    - predikce
    - detekce anomálií
  - graph
    - analýza vazeb (vztahů) mezi daty
    - plugin do Kibany pro vizualizaci vazeb
  - Elasticsearch SQL
    - dotazování pomocí SQL
    - dotazy jsou běžně psány v Query DSL
    - překlad SQL do Query DSL na pozadí

https://www.udemy.com/course/elasticsearch-complete-guide/

BEATS

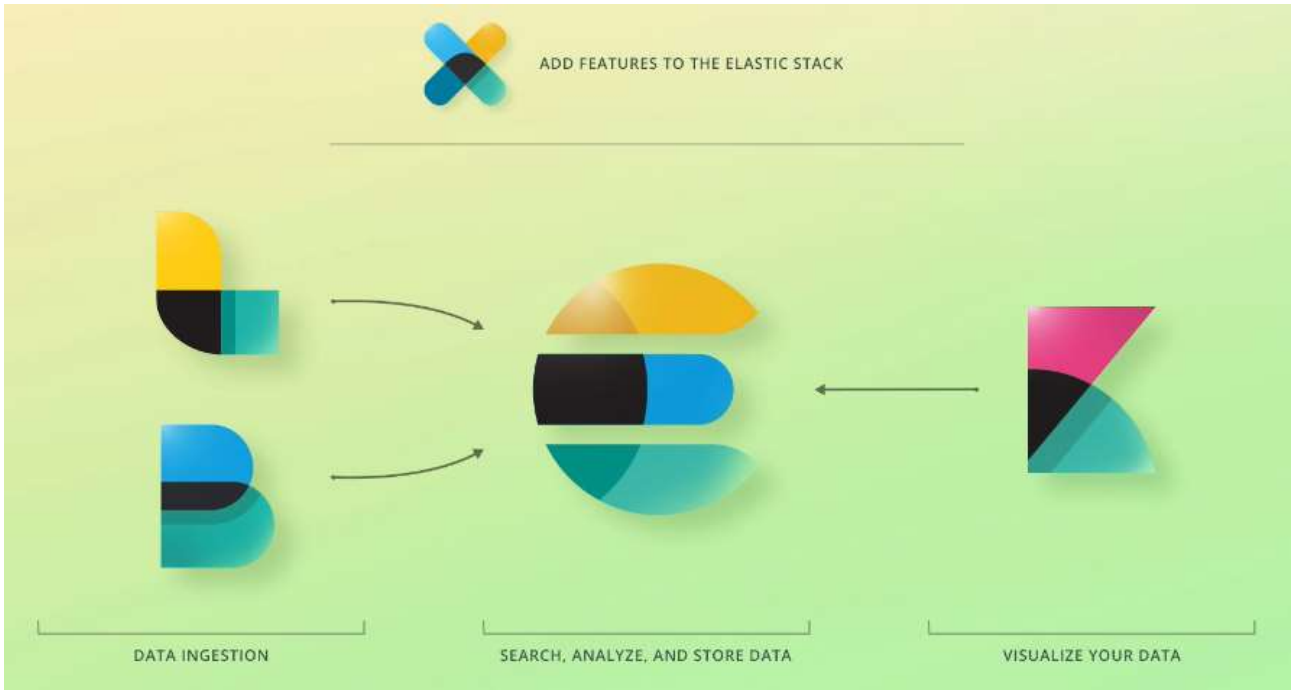
- kolekce data shippers
  - nenároční agenti instalovaní na servery
  - posílají data do Logstash / Elasticsearch
  - různé druhy
    - sběr různých dat
      - filebeat – sběr logů
      - metricbeat – systémové metriky





https://www.udemy.com/course/elasticsearch-complete-guide/

ELASTIC STACK



https://www.udemy.com/course/elasticsearch-complete-guide/

A PŘÍŠTĚ?

- sloupcové databáze
  - Apache Cassandra

Děkuji za pozornost.
Otázky?