Florian Hopf - @fhopf

GOTO nights Berlin
22.10.2015

Data modeling for

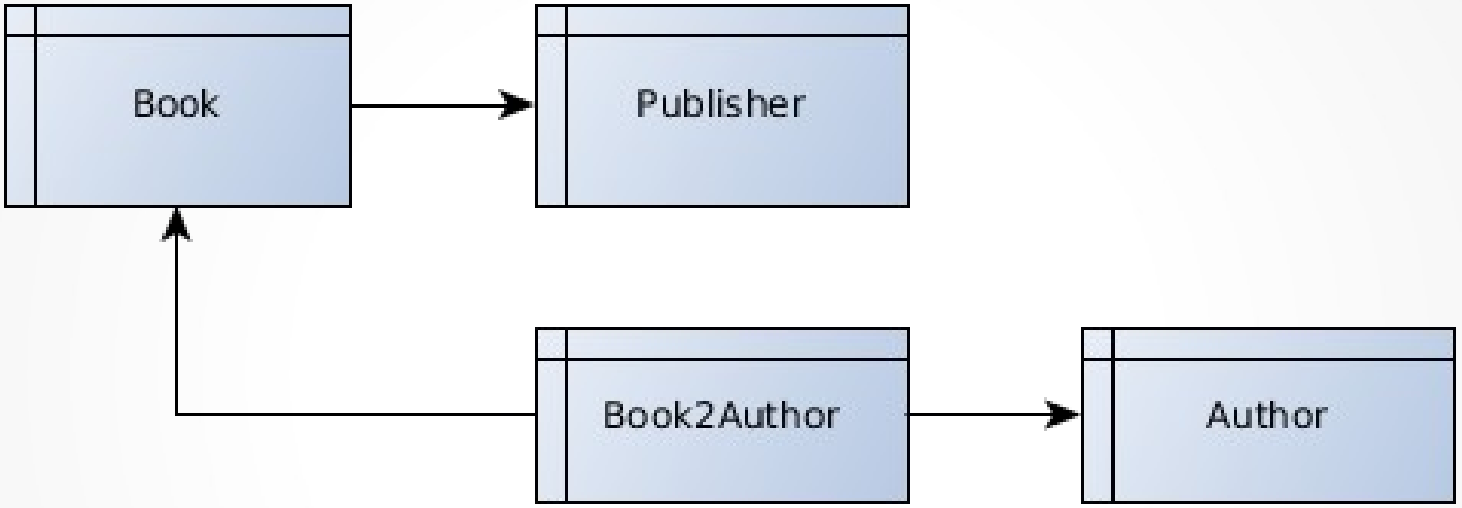
elasticsearch.

What are we talking about?

- Storing and querying data
  - String
  - Numeric
  - Date
  - Embedding documents
- Types and Mapping
- Updating data
- Time stamped data

Documents

A relational view



A relational view

- Different aspects are stored in different tables
- Traversal of tables via join-Operations
- High degree of normalization

Documents

{
Book
Author
Publisher
}

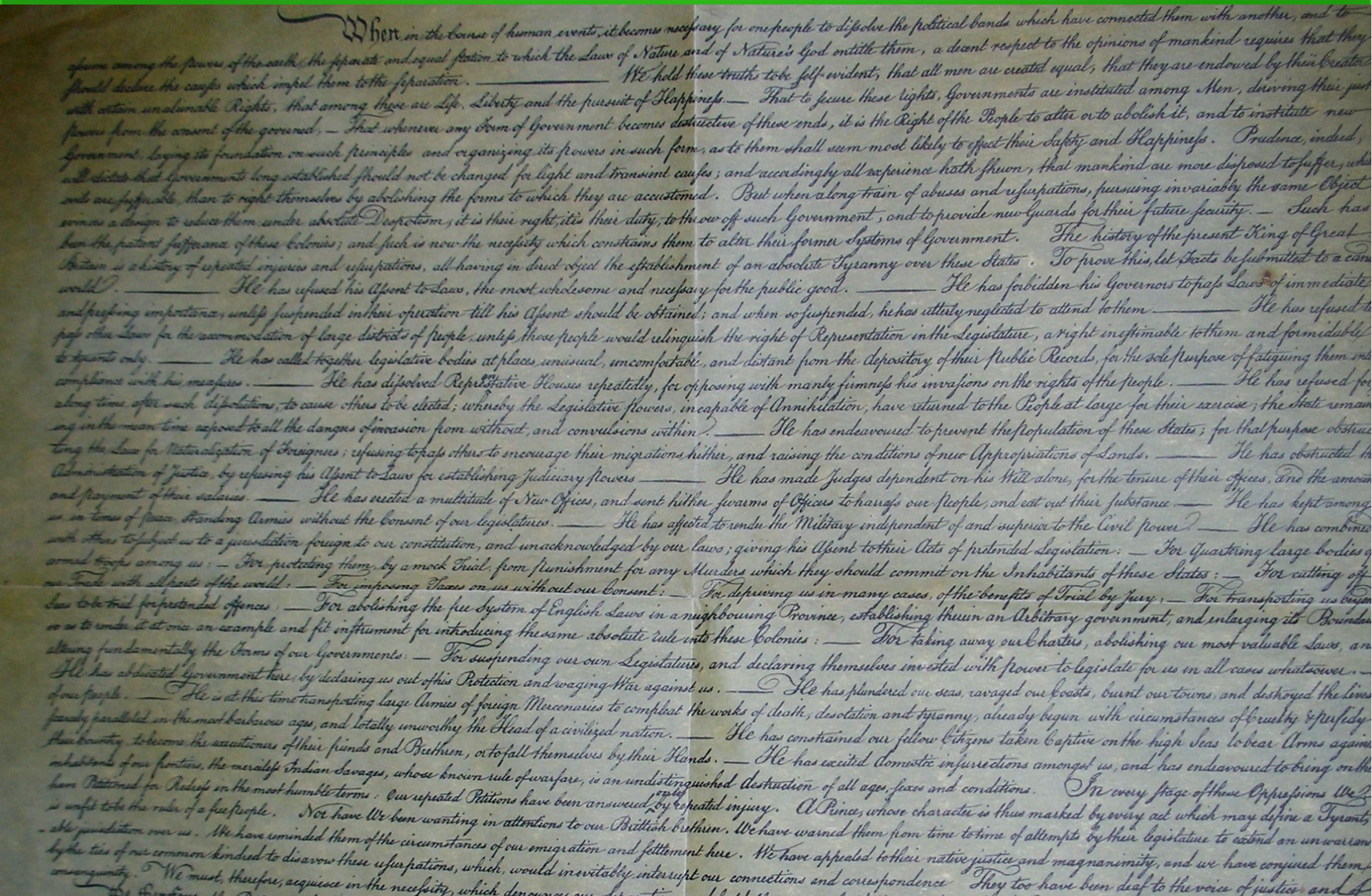
Documents

- Often more natural
- Flexible schema
- Fields can be queried
- Duplicate storage of document parts

Documents

```
POST /library/book
{
    "title": "Elasticsearch in Action",
    "author": [ "Radu Gheorghe",
                "Matthew Lee Hinman",
                "Roy Russo" ],
    "pages": 400,
    "published": "2015-06-30T00:00:00.000Z",
    "publisher": {
        "name": "Manning",
        "country": "USA"
    }
}
```

Text



When in the course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to
assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they
should declare the causes which impel them to the separation. ——— We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator
with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. — That to secure these rights, Governments are instituted among Men, deriving their just
powers from the consent of the governed, — That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new
Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed,
will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn, that mankind are more disposed to suffer, while
evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, pursuing invariably the same Object
evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security. — Such has
been the patient sufferance of these Colonies; and such is now the necessity which constrains them to alter their former Systems of Government. The history of the present King of Great
Britain is a history of repeated injuries and usurpations, all having in direct object the establishment of an absolute Tyranny over these States. To prove this, let Facts be submitted to a can-
did world. ——— He has refused his Assent to Laws, the most wholesome and necessary for the public good. ——— He has forbidden his Governors to pass Laws of immediate
and pressing importance, unless suspended in their operation till his Assent should be obtained; and when so suspended, he has utterly neglected to attend to them ——— He has refused to
pass other Laws for the accommodation of large districts of people, unless those people would relinquish the right of Representation in the Legislature, a right inestimable to them and formidable
to tyrants only. ——— He has called together legislative bodies at places unusual, uncomfortable, and distant from the depository of their public Records, for the sole purpose of fatiguing them into
compliance with his measures. ——— He has dissolved Representative Houses repeatedly, for opposing with manly firmness his invasions on the rights of the people. ——— He has refused for
a long time, after such dissolutions, to cause others to be elected; whereby the Legislative powers, incapable of Annihilation, have returned to the People at large for their exercise; the State remain-
ing in the mean time exposed to all the dangers of invasion from without, and convulsions within. ——— He has endeavoured to prevent the population of these States; for that purpose obstruct-
ing the Laws for Naturalization of Foreigners; refusing to pass others to encourage their migrations hither, and raising the conditions of new Appropriations of Lands. ——— He has obstructed the
Administration of Justice, by refusing his Assent to Laws for establishing Judiciary powers ——— He has made Judges dependent on his Will alone, for the tenure of their offices, and the amount
and payment of their salaries. ——— He has erected a multitude of New Offices, and sent hither swarms of Officers to harass our people, and eat out their substance. ——— He has kept among
us, in times of peace, Standing Armies without the Consent of our legislatures. ——— He has affected to render the Military independent of and superior to the Civil power. ——— He has combined
with others to subject us to a jurisdiction foreign to our constitution, and unacknowledged by our laws; giving his Assent to their Acts of pretended Legislation: — For Quartering large bodies of
armed troops among us: — For protecting them, by a mock Trial, from punishment for any Murders which they should commit on the Inhabitants of these States: — For cutting off
our Trade with all parts of the world: — For imposing Taxes on us without our Consent: — For depriving us in many cases, of the benefits of Trial by Jury: — For transporting us beyond
seas to be tried for pretended offences: — For abolishing the free System of English Laws in a neighbouring Province, establishing therein an Arbitrary government, and enlarging its Boundaries
so as to render it at once an example and fit instrument for introducing the same absolute rule into these Colonies: — For taking away our Charters, abolishing our most valuable Laws, and
altering fundamentally the Forms of our Governments: — For suspending our own Legislatures, and declaring themselves invested with power to legislate for us in all cases whatsoever. —
He has abdicated Government here, by declaring us out of his Protection and waging War against us. ——— He has plundered our seas, ravaged our Coasts, burnt our towns, and destroyed the lives
of our people. ——— He is at this time transporting large Armies of foreign Mercenaries to compleat the works of death, desolation and tyranny, already begun with circumstances of Cruelty & Perfidy
scarcely paralleled in the most barbarous ages, and totally unworthy the Head of a civilized nation. ——— He has constrained our fellow Citizens taken Captive on the high Seas to bear Arms against
their Country, to become the executioners of their friends and Brethren, or to fall themselves by their Hands. ——— He has excited domestic insurrections amongst us, and has endeavoured to bring on the
inhabitants of our frontiers, the merciless Indian Savages, whose known rule of warfare, is an undistinguished destruction of all ages, sexes and conditions. In every stage of these Oppressions We
have Petitioned for Redress in the most humble terms: Our repeated Petitions have been answered only by repeated injury. A Prince, whose character is thus marked by every act which may define a Tyrant,
is unfit to be the ruler of a free people. Nor have We been wanting in attentions to our British brethren. We have warned them from time to time of attempts by their legislature to extend an unwarrant-
able jurisdiction over us. We have reminded them of the circumstances of our emigration and settlement here. We have appealed to their native justice and magnanimity, and we have conjured them
by the ties of our common kindred to disavow these usurpations, which, would inevitably interrupt our connections and correspondence. They too have been deaf to the voice of justice and of
consanguinity. We must, therefore, acquiesce in the necessity, which denounces our

Text

```
POST /library/book
{
    "title": "Elasticsearch in Action",
    "author": [ "Radu Gheorghe",
                "Matthew Lee Hinman",
                "Roy Russo" ],
    "pages": 400,
    "published": "2015-06-30T00:00:00.000Z",
    "publisher": {
        "name": "Manning",
        "country": "USA"
    }
}
```

Searching data

```
GET /library/book/_search?q=elasticsearch

{
   "took": 75,
   "timed_out": false,
   "_shards": {
      "total": 5,
      "successful": 5,
      "failed": 0
   },
   "hits": {
      "total": 1,
      "max_score": 0.067124054,
      "hits": [
        [...]
      ]
   }
}
```

Searching data

```
GET /library/book/_search
{
    "query": {
        "match": {
           "title": "elasticsearch"
        }
    }
}
```

Understand index storage

- Data is stored in the inverted index
- Analyzing process determines storage and query characteristics
- Important for designing data storage

Analyzing

Elasticsearch in Action

Elasticsearch: Ein praktischer Einstieg

1. Tokenization

Term	Document Id
Action	1
ein	2
Einstieg	2
Elasticsearch	1,2
in	1
praktischer	2

Analyzing

Elasticsearch in Action

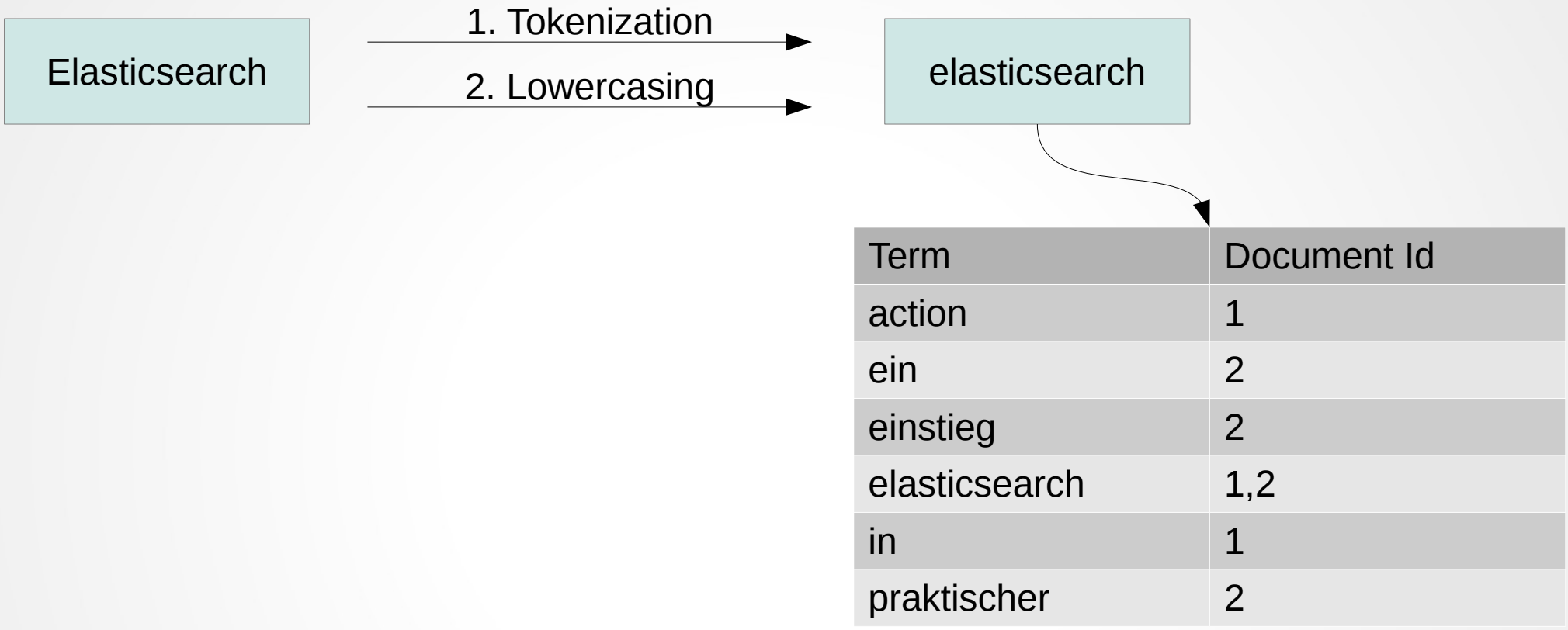
Elasticsearch: Ein praktischer Einstieg

1. Tokenization

2. Lowercasing

Term	Document Id
action	1
ein	2
einstieg	2
elasticsearch	1,2
in	1
praktischer	2

Search



Term	Document Id
action	1
ein	2
einstieg	2
elasticsearch	1,2
in	1
praktischer	2

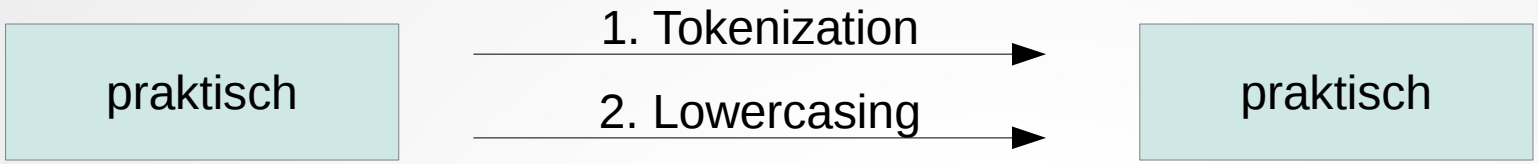
Inverted Index

- Terms are deduplicated
- Original content is lost
- Elasticsearch stores the original content in a special field source

Inverted Index

- New requirement: search for German content
  - praktischer → praktisch

Search



Term	Document Id
action	1
ein	2
einstieg	2
elasticsearch	1,2
in	1
praktischer	2

Analyzing

Elasticsearch in Action

Elasticsearch: Ein praktischer Einstieg

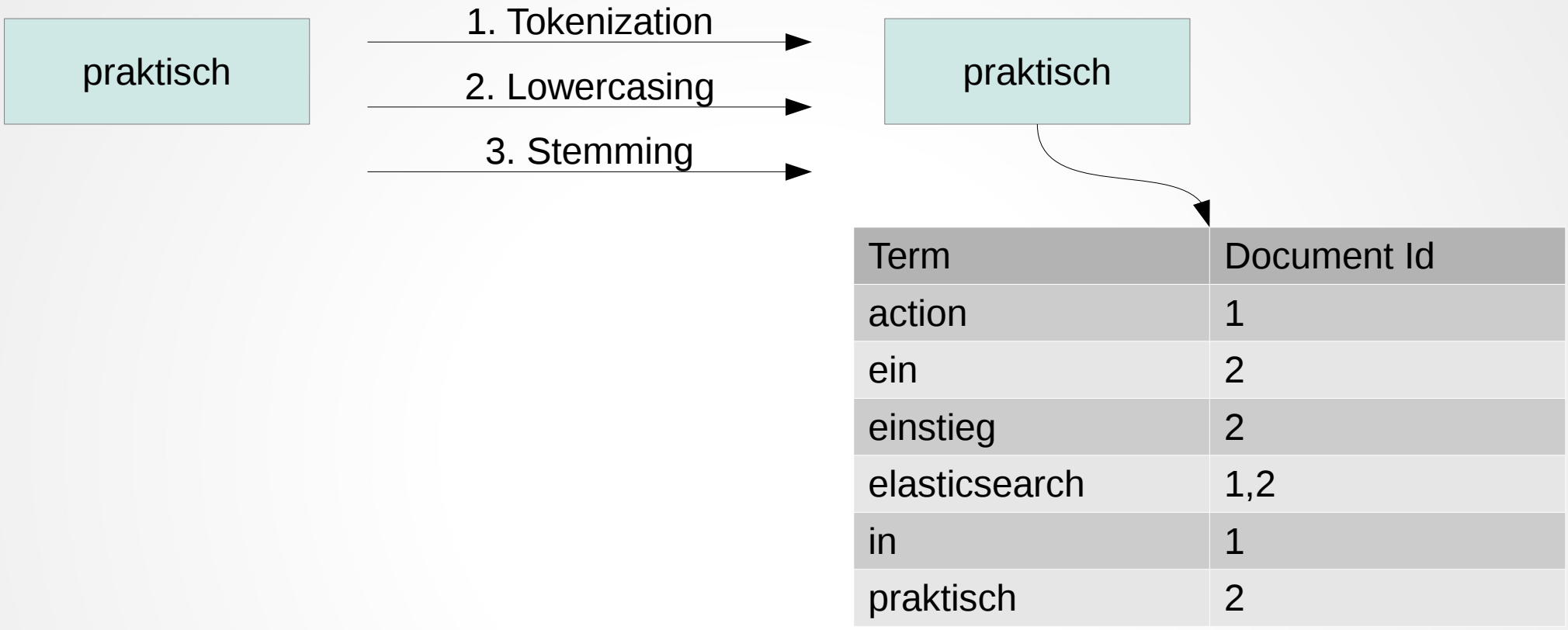
1. Tokenization

2. Lowercasing

3. Stemming

Term	Document Id
action	1
ein	2
einstieg	2
elasticsearch	1,2
in	1
praktisch	2

Search



Term	Document Id
action	1
ein	2
einstieg	2
elasticsearch	1,2
in	1
praktisch	2

Mapping

```
curl -XPUT "http://localhost:9200/library/book/_mapping"
-d'
{
  "book": {
    "properties": {
      "title": {
        "type": "string",
        "analyzer": "german"
      }
    }
  }
}'
```

Understand index storage

- For every indexed document Elasticsearch builds a mapping from the fields in the documents
- Sane defaults for lots of use cases
- But: understand and control it and your data

Searching data

```
GET /library/book/_search?q=elasticsearch

{
   "took": 75,
   "timed_out": false,
   "_shards": {
      "total": 5,
      "successful": 5,
      "failed": 0
   },
   "hits": {
      "total": 1,
      "max_score": 0.067124054,
      "hits": [
        [...]
      ]
   }
}
```

_all

- Default search field _all

```
"book": {
   "_all": {
      "enabled": false
   }
}
```

Partial Word Matches

- New requirement: Search for parts of words
  - elastic → elasticsearch

Partial Word Matches

- Common option: Using wildcards

```
POST /library/book/_search
{
    "query": {
        "wildcard": {
            "title": {
                "value": "elastic*"
            }
        }
    }
}
```

Partial Word Matches

- Wildcards
  - Query time option
  - Scalability?

Partial Word Matches

- Alternative: Index Time preprocessing
  - Terms are stored in the index in a special way
  - Search is then a normal lookup
  - For partial words: N-Grams

N-Grams

- Configuring an N-Gram analyzer
- Builds N-Grams
  - elas
  - elast
  - elasti
  - elastic
  - elastics
  - ...

Index Settings for N-Grams

```
PUT /library-ngram
{
  "settings": {
    "analysis": {
      "analyzer": {
        "prefix_analyzer": {
          "type": "custom",
          "tokenizer": "prefix_tokenizer",
          "filter": ["lowercase"]
        }
      },
      "tokenizer": {
        "prefix_tokenizer": {
          "type": "edgeNGram",
          "min_gram" : "4",
          "max_gram" : "8",
          "token_chars": [ "letter", "digit" ]
        }
      }
    }}}
```

Mapping for N-Grams

```
PUT /library-ngram/book/_mapping
{
  "book": {
    "properties": {
      "title": {
        "type": "string",
        "analyzer": "german",
        "fields": {
          "prefix": {
            "type": "string",
            "index_analyzer": "prefix_analyzer",
            "query_analyzer": "lowercase"
          }
        }
      }
    }
  }
}
```

Additional Field

- Indexed Document stays the same
- Additional index field title.prefix
- Can be queried like any field

Querying additional Field

```
GET /library-ngram/book/_search
{
    "query": {
        "match": {
           "title.prefix": "elastic"
        }
    }
}
```

Querying additional Field

```
GET /library-ngram/book/_search
{
   "query": {
      "bool": {
         "should": [
            {
               "match": {
                  "title": "elastic"
               }
            },
            {
               "match": {
                  "title.prefix": "elastic"
               }
            }
         ]
      }
   }
}
```

Additional Field

- Increased storage requirements
- Increased scalability (and performance) during search
- Trade storage against search performance

Numbers


BINGO
1 2 3 4 5 6 7 8 9 10
16 17 18 19 20 21 22 23 24
31 32 33 34 35 36 37 38 39
46 47 48 49 50 51 52 53 54
61 62 63 64 65 66 67

Storing data

```
POST /library/book
{
    "title": "Elasticsearch in Action",
    "author": [ "Radu Gheorghe",
                "Matthew Lee Hinman",
                "Roy Russo" ],
    "pages": 400,
    "published": "2015-06-30T00:00:00.000Z",
    "publisher": {
        "name": "Manning",
        "country": "USA"
    }
}
```

Querying

- Numeric term is in index

```
POST /library/book/_search
{
    "query": {
        "term": {
            "pages": "400"
        }
    }
}
```

Querying

- Ranges

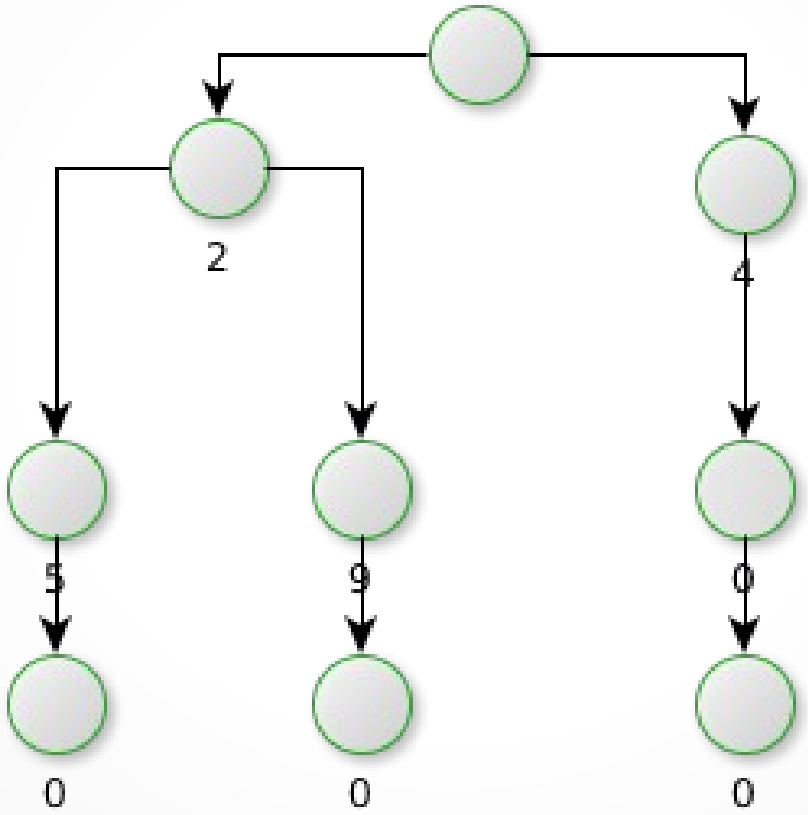
```
POST /library/book/_search
{
    "query": {
        "range": {
            "pages": {
                "gte": 300
            }
        }
    }
}
```

Numeric values

- Numeric values are stored in a Trie structure
- Makes range queries very efficient

Numeric values

- Simplified view: 250, 290 and 400



Numeric values

- Precision influences depth of tree
- Lower precision_step → higher number of terms
- Most of the time defaults are fine

Date

Storing data

```
POST /library/book
{
    "title": "Elasticsearch in Action",
    "author": [ "Radu Gheorghe",
                "Matthew Lee Hinman",
                "Roy Russo" ],
    "pages": 400,
    "published": "2015-06-30T00:00:00.000Z",
    "publisher": {
        "name": "Manning",
        "country": "USA"
    }
}
```

Date

- Default: ISO8601 format
- Joda Time patterns
- Internally stored as long

Date

```
PUT /library-date/book/_mapping
{
    "book": {
        "properties": {
            "published": {
                "type": "date",
                "format": "dd.MM.yyyy"
            }
        }
    }
}
```

Date

```
POST /library-date/book
{
    "title": "Elasticsearch in Action",
    "author": [ "Radu Gheorghe",
           "Matthew Lee Hinman",
           "Roy Russo" ],
    "pages": 400,
    "published": "30.06.2015",
    "publisher": {
        "name": "Manning",
        "country": "USA"
    }
}
```

Date

- Common: Filtering on date range
- from and/or to

Date

```
"query": {
    "filtered": {
        "filter": {
            "range": {
                "published": {
                    "to": "30.06.2015"
                }
            }
        }
    }
}
```

Date

```
"query": {
    "filtered": {
       "filter": {
           "range": {
              "published": {
                 "to": "now-3M"
              }
           }
       }
    }
}
```

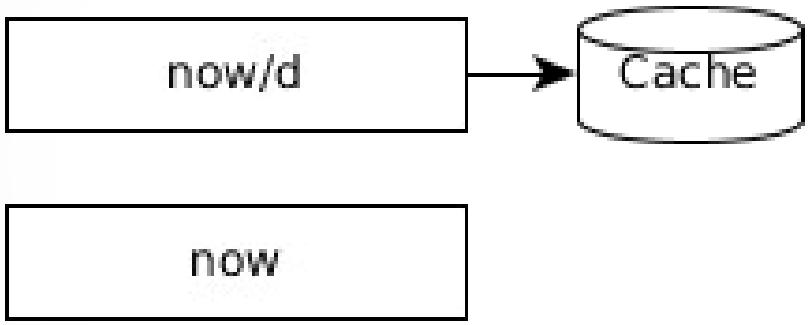
Date

- Filter is not cached with 'now'
- Only cached with rounded value

```
"range": {
   "published": {
      "to": "now-3M/d"
   }
}
```

Date

- Exact values needed → Combine filters



Embedded Documents

Embedded Documents

```
POST /library/book
{
    "title": "Elasticsearch in Action",
    "author": [ "Radu Gheorghe",
                "Matthew Lee Hinman",
                "Roy Russo" ],
    "pages": 400,
    "published": "2015-06-30T00:00:00.000Z",
    "publisher": {
        "name": "Manning",
        "country": "USA"
    }
}
```

Embedded Documents

- Default: Flat structure
- Good for 1:1 relation

```
"publisher": {
        "name": "Manning",
        "country": "USA"
}
```

```
"publisher.name": "Manning",
"publisher.country": "USA"
```

Embedded documents

- 1:N relations are problematic

```
{
    "title": "Elasticsearch in Action",
    "ratings": [
    {
        "source": "Amazon",
        "stars": 5
    },
    {
        "source": "Goodreads",
        "stars": 4
    }
    ]
}
```

Embedded documents

- 1:N relations are problematic

```
"query": {
   "bool": {
      "must": [
         { "match": { "ratings.source": "Goodreads" }},
         { "match": { "ratings.stars": 5 }}
      ]
   }
}
```

Nested

- Solution: Nested documents
- Lucene internal: Seperate document, connected via Block-Join
- Accessing documents via specialized query

Nested

- Explicit mapping

```
"book": {
   "properties": {
      "ratings": {
         "type": "nested",
         "properties": {
            "source": {
               "type": "string"
            },
            "stars": {
               "type": "integer"
            }
         }
      }
   }
}
```

Nested

- Nested-Query

```
"query": {
  "nested": {
    "path": "ratings",
    "query": {
      "bool": {
        "must": [
          { "match": { "ratings.source": "Goodreads" }},
          { "match": { "ratings.stars": 5 }}
        ]
      }
    }
  }
}
```

Nested

- Additional flat storage
  - include_in_parent
  - include_in_root

Parent-Child

- Alternative storage
- Indexing seperate types
- Connection via parent parameter

Parent-Child

- Book is stored without ratings

```
POST /library-parent-child/book/
{
    "title": "Elasticsearch in Action",
    "publisher": {
        "name": "Manning"
    }
}
```

Parent-Child

- Ratings reference books

```
PUT /library-parent-child/rating/_mapping
{
  "rating": {
    "_parent": {
      "type": "book"
    }
  }
}
```

Parent-Child

- Ratings reference book

```
POST /library-parent-child/rating?
parent=AU_smK5FYK634dNiekGr
{
  "source": "Amazon",
  "stars": 5
}

POST /library-parent-child/rating?
parent=AU_smK5FYK634dNiekGr
{
  "source": "Goodreads",
  "stars": 4
}
```

Parent-Child

- has_child/has_parent

```
POST /library-parent-child/book/_search
{
  "query": {
    "has_child": {
      "type": "rating",
      "query": {
        "bool": {
           "must": [
              { "match": {"source": "Goodreads" }},
              { "match": {"stars": 5 }}
           ]
        }
      }
    }
  }
}
```

Parent-Child

- Stored on same shard
- Only suitable for smaller amounts of docs
- Requires different types

Types and Mapping

Querying Elasticsearch

- Ad-hoc queries
  - But better characteristics when designing storage for query
- Flexible Schema
  - But mapping better defined upfront

Mapping

- Mapping for field can't be changed
- Think about how you will be querying your data
- Think about defining a static mapping upfront

Disable dynamic mapping

```
PUT /library/book/_mapping
{
    "book": {
        "dynamic": "strict"
    }
}
```

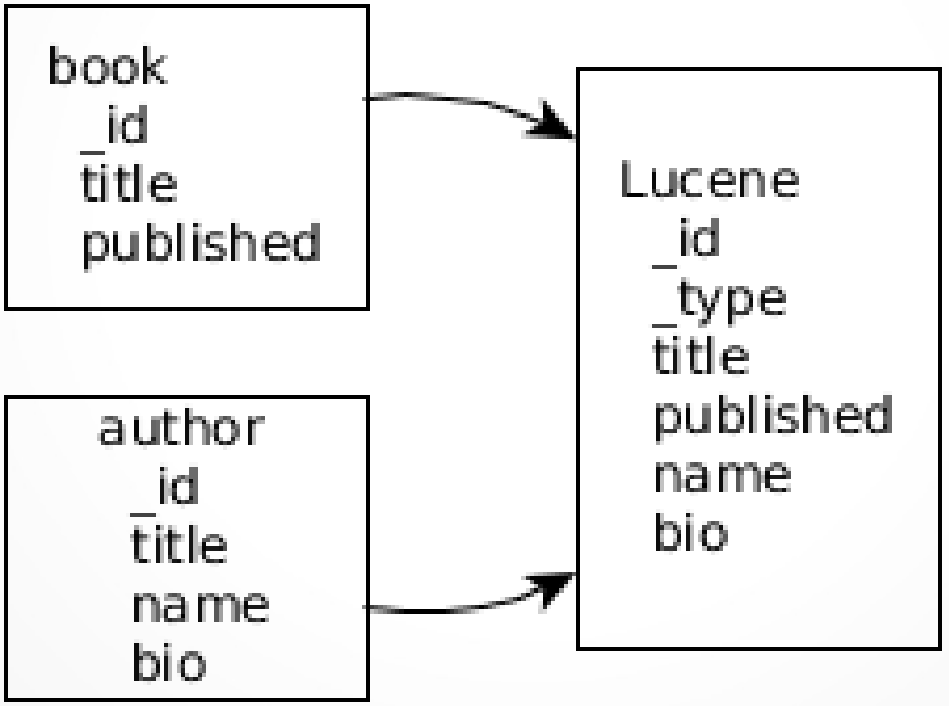
Disable dynamic mapping

```
POST /library/book
{
    "titel": "Falsch"
}

{
"error" : "StrictDynamicMappingException[mapping set to
strict! dynamic introduction of [titel] within [book]
is not allowed]",
"status" : 400
}
```

Types

- Types determine mapping
- Lucene doesn't know about types



Types

- Fields with same names need to be mapped the same way
- Relevance can be influenced
- Index settings: shards, replicas per type?

Key-Value-Store

- Careful when using ES as key-value-store
- Mapping is part of cluster state

Updating Data

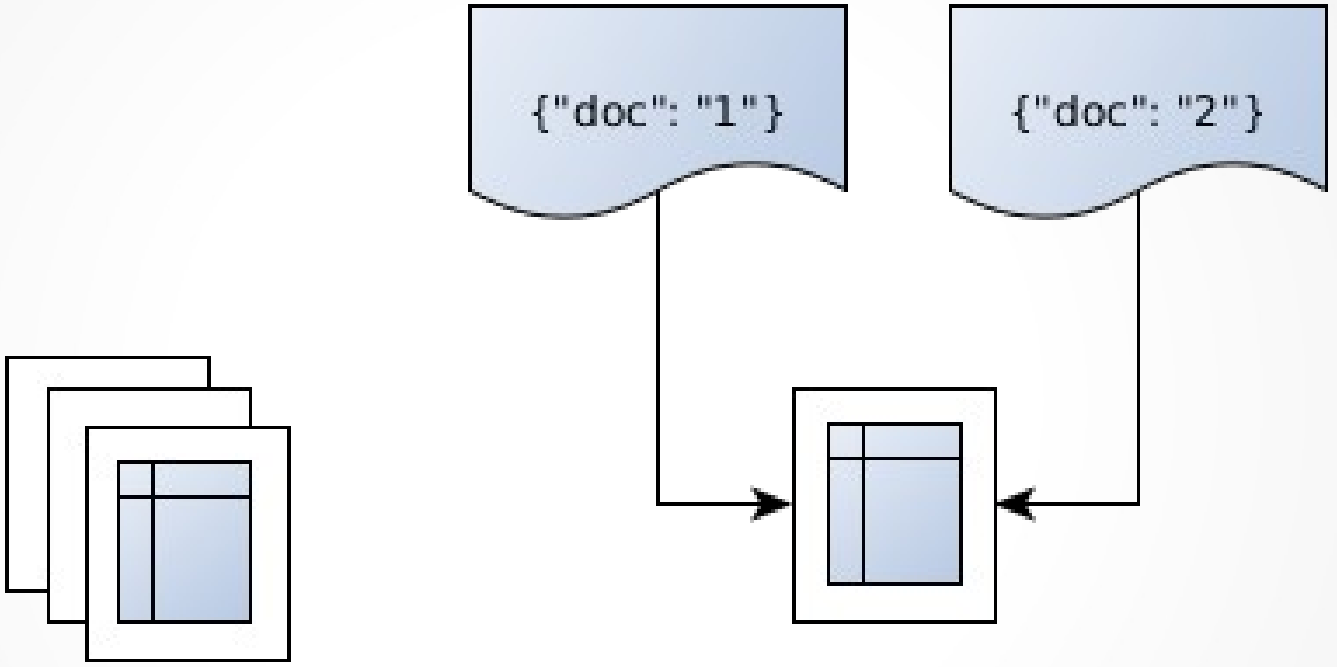
Updating Data

- Primary Datastore
- Full indexing
- Incremental indexing

Updating Data

- Elasticsearch stores data in segment files
- Immutable files
- Segment is a mini inverted index

Segments

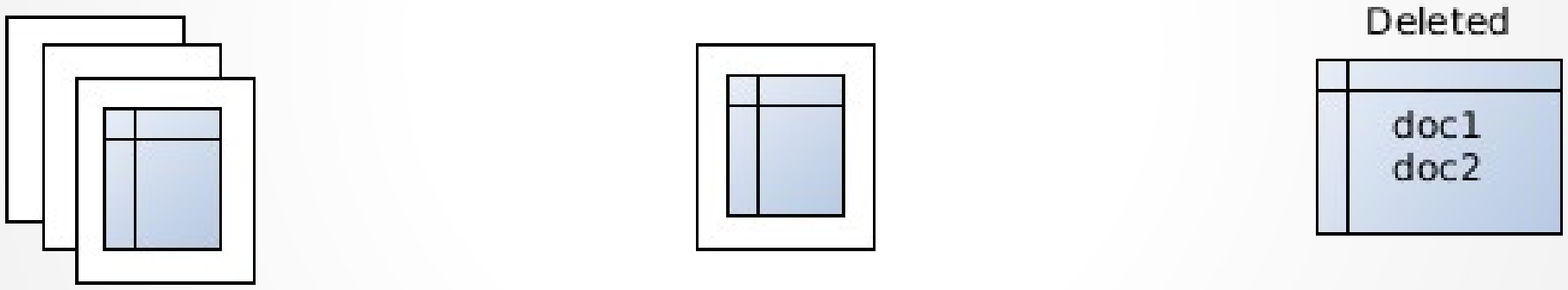


Segments

- Building inverted index is expensive
- Add documents → add new segments

Segments

- Doc deletion is only a marker
- Deleted documents are automatically filtered



Updating Data

- Documents can be updated
- Full Update
- Partial Update

Updating data

- Full update: Replaces a document

```
PUT /library/book/AVBDusjh0tduyhTzZqTC
{
   "title": "Elasticsearch in Action",
   "author": [
      "Radu Gheorghe",
      "Matthew L. Hinman",
      "Roy Russo"
   ],
   "published": "2015-06-30T00:00:00.000Z",
   "publisher": {
      "name": "Manning",
      "country": "USA"
   }
}
```

Updating data

- Partial update: Uses source of document

```
POST /library/book/AVBDusjh0tduyhTzZqTC/_update
{
   "doc": {
      "title": "Elasticsearch In Action"
   }
}
```

Updating data

- Update = Delete + Add
- Expensive operation
- Design documents as events if possible

Timestamps

Working with timestamps

- Timestamped data
- Write events
- Common: Log events

Index Design

- Use date aware index name
- library-221015
- Create a new index every day

Index Design

- Index templates for custom settings

```
PUT /_template/library-template
{
   "template": "library-*",
   "mappings": {
      "book": {
         "properties": {
            "title": {
               "type": "string",
               "analyzer": "german"
            }
         }
      }
   }
}
```

Index Design

- Search multiple indices

```
GET /library-221015,library-211015/_search

GET /library-*/_search
```

Index Design

- Combining indices with Index-Aliases

```
POST /_aliases
{
    "actions" : [
        { "add" : {
           "index" : "library-2015*",
           "alias" : "thisyear"
        }},
        { "add" : {
           "index" : "library-2015-10*",
              "alias" : "thismonth"
        }}
    ]
}
```

Index Design

- Implicit date selection

```
GET /thisyear/_search

GET /thismonth/_search
```

Index Design

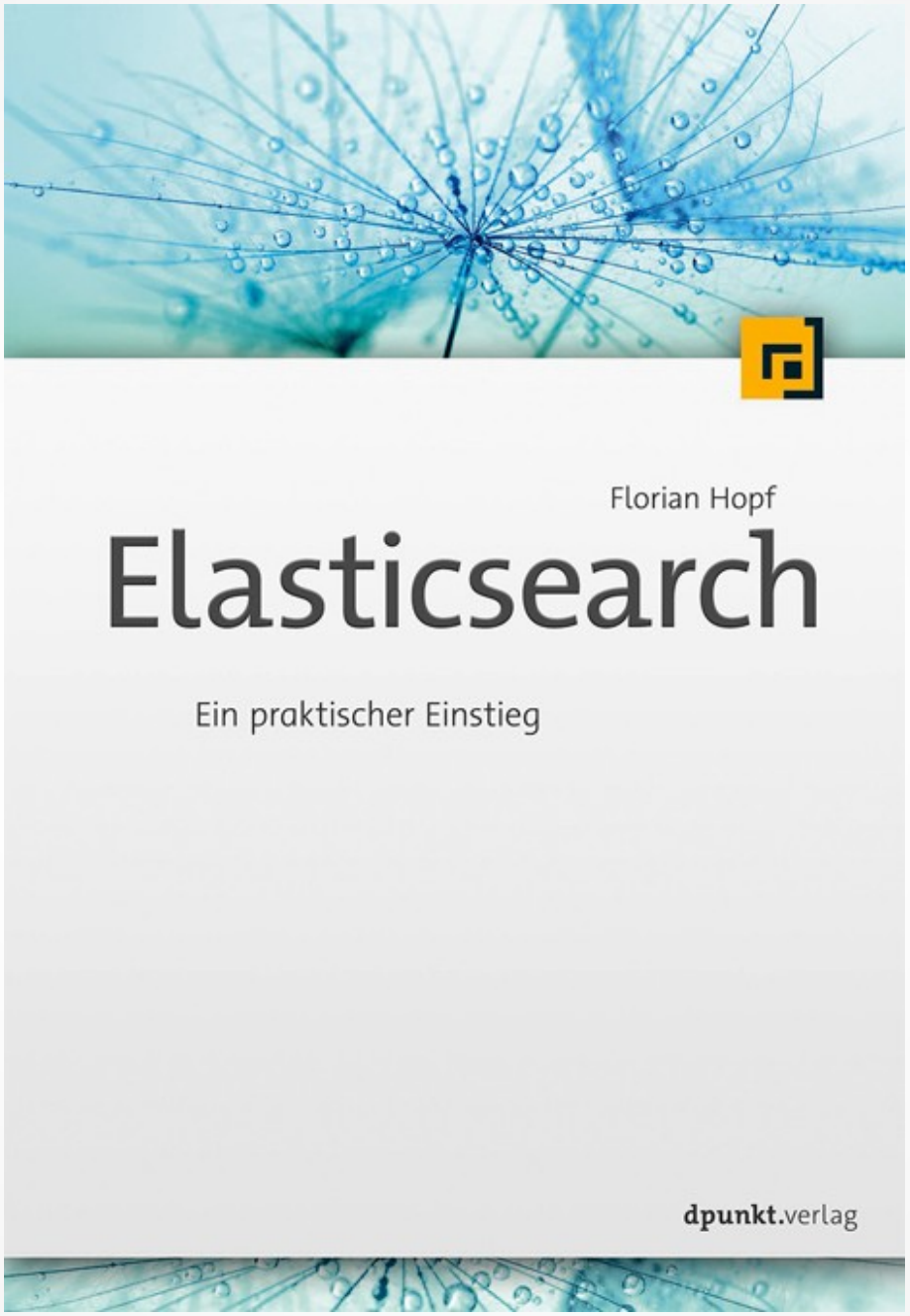
- Filtered Alias

```
"actions" : [{
   "add" : {
      "index" : "library",
      "alias" : "buecher",
      "filter" : {
         "term" : { "publisher.country" : "de" }
      }
   }
}]
```

What is missing?

- Distributed data and Routing
- Field Data and Doc Values
- Index-Options
- Geo-Data

More Info


Florian Hopf
Elasticsearch
Ein praktischer Einstieg
dpunkt.verlag

More Info

- http://elastic.co
- Elasticsearch – The definitive Guide
  - https://www.elastic.co/guide/en/elasticsearch/guide/master/index.html
- Elasticsearch in Action
  - https://www.manning.com/books/elasticsearch-in-action
- http://blog.florian-hopf.de

Resources

- http://blog.parsely.com/post/1691/lucene/
- http://de.slideshare.net/VadimKirilchuk/numeric-rangequeries
- https://www.elastic.co/blog/found-optimizing-elasticsearch-searches

Images

- http://www.morguefile.com/archive/display/48456
- http://www.morguefile.com/archive/display/104082
- http://www.morguefile.com/archive/display/978102
- http://www.morguefile.com/archive/display/978102
- http://www.morguefile.com/archive/display/861633
- http://www.morguefile.com/archive/display/899572
- http://www.morguefile.com/archive/display/903066
- http://www.morguefile.com/archive/display/53012