

# Search Algorithms and Data Structures

Mikhail Khudnev

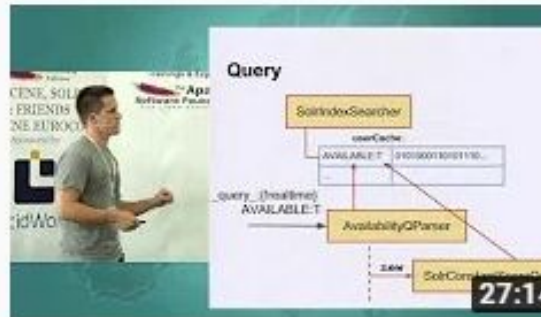
Khabarovsk'19

# About Me

Search or enter web l

Mikhail

FILTER



Indices: 3,705

Documents

87,310,395,558

Disk Usage

154.3 TB

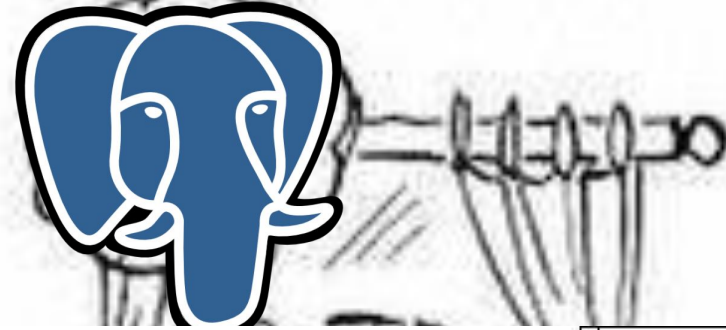
Primary Shards

8,222

Replica Shards

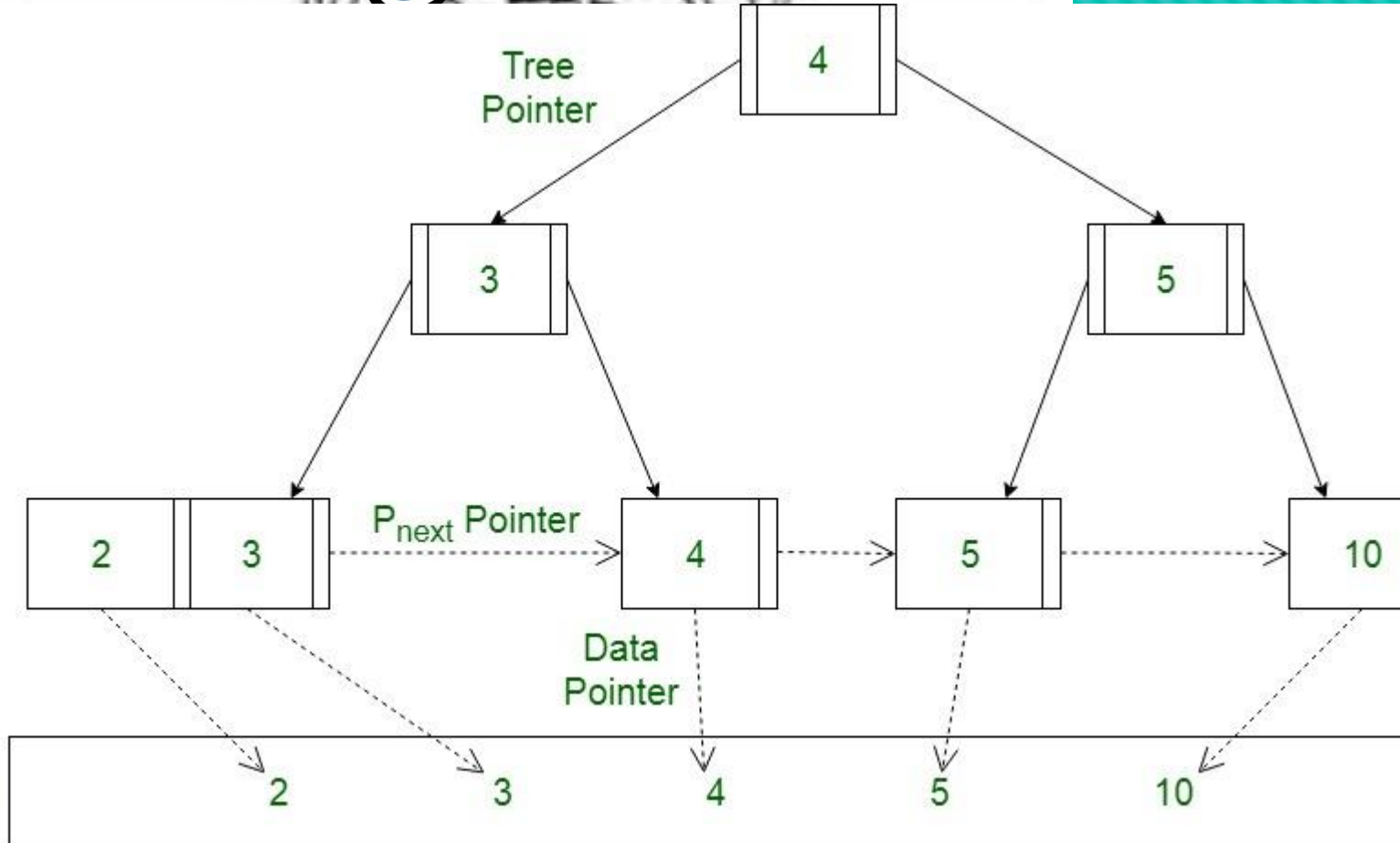
13,124

# RDBMS is ...



Outcome					
ID	ESNum	Type	TxN	CgN	ES
100	1	1	24	24	-0.39
100	2	1	24	24	0
100	3	1	24	24	0.09
100	4	1	24	24	-1.05
100	5	1	24	24	-0.44
7049	1	2	30	30	0.34
7049	2	4	30	30	0.78
7049	3	1	30	30	0

Disk File



<https://www.geeksforgeeks.org/database-file-indexing-b-tree-introduction/>

# Composite Index

```
CREATE INDEX class_pos_index ON users (class, position);
```

## Size

### Heel Height

- Low 1-2" (120)
- Mid 2-3" (399)
- High 3-4" (415)
- Ultra High 4" & Over (157)

### Heel Type

- Block (306)
- Cone (16)
- Kitten (117)
- Sculpted (10)
- Stacked (152)
- Stiletto (352)
- Wedge (42)

## Brand

🔍 Search by brand

- Adrianna Papell (17)
- Adrienne Vittadini (8)
- Aerosoles (14)

896 items in Heels & Pumps

Free Pickup Today [Macy's Jersey City](#)



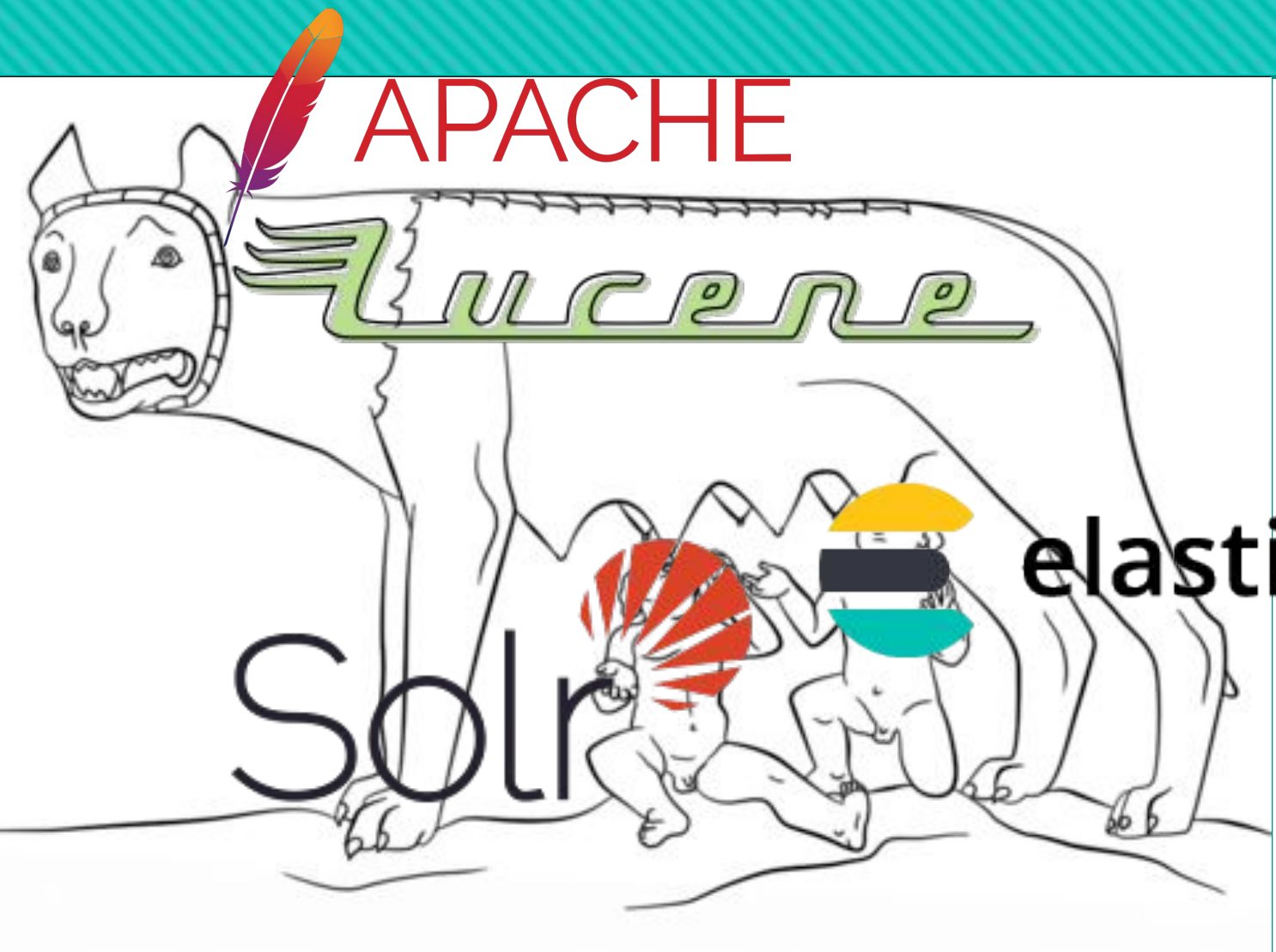
MICHAEL Michael Kors Berkley T-Strap Platform

Extended Widths



# Inverted Index

# Lucene, Solr and Elasticsearch



<http://www.supercoloring.com/coloring-pages/remulus-and-remus-with-the-she-wolf>

# The First Indices

¶ Tabula brevis et utilis super libello  
quodam qui dicitur fasciculus temporum:  
et ubi inuenitur punctus ante numerum est  
in primo latere: ubi vero post in secundo  
latere: Incipit feliciter.

Abacū p̄ba minor	15
Abacū martyr	32
Abdias p̄ba	11 12
Abdon iudex	8
Abdon et sennos martyres	32
Abdo p̄ba extra hierosolā	9
Abel p̄m martyr innocēs	2
Abellan iudex isrl̄ton	8
Abya rex iuda	10
Abiabar pontifex ḡstosus	9
Abu filius aaron	7
Abymelech iudex isrl̄ malus	8
Abymelech pontifex	10
Abisus pontifex	8
Abiathar rex isrl̄ corobabel	17
Abzahā nascif inchoans iudā et arā	5
Abraham moisi	6
A c	
¶ Achaz rex iuda	13
Achtold' ep̄s vintoniē.	52
Achias filionis p̄ba	9
Achymas fil' sadoch pontifex	9
Abymelech sacerdos dñi	9
Achim rex	19
Achis filius rex alban'	9
Achitob pontifex	9
Achitob sacerdos	9
Achitob pontifex	13
Acies ignee sup hierosolā	21
Acies ignee visē sūt in celo	40
Aciō ciuitas fortissima capis	59
A d	
¶ Ada uel oda vxor lamech	3
Ado p̄ba	10
Adam p̄mus hō formaf	2
Adam moisi	3
Adamaras vir sc̄us	51
Additio annor' egechie regis	13
Adcodatus papa	43
Adelternus vir sc̄us	52
Adelternus ep̄s metes.	53
Ado vir religiosus	42
Adolppus impator	60
Adrian' h̄xiē impator	29
Adrianus papa p̄mus	47
Adrianus papa secund'	49
Adrianus papa tert'	49
Adrian' papa quart'	56
Adrian' papa quint'	59
Adrian' cronob' et mōch'	54
A f	
¶ African' sc̄ipio roman'	21

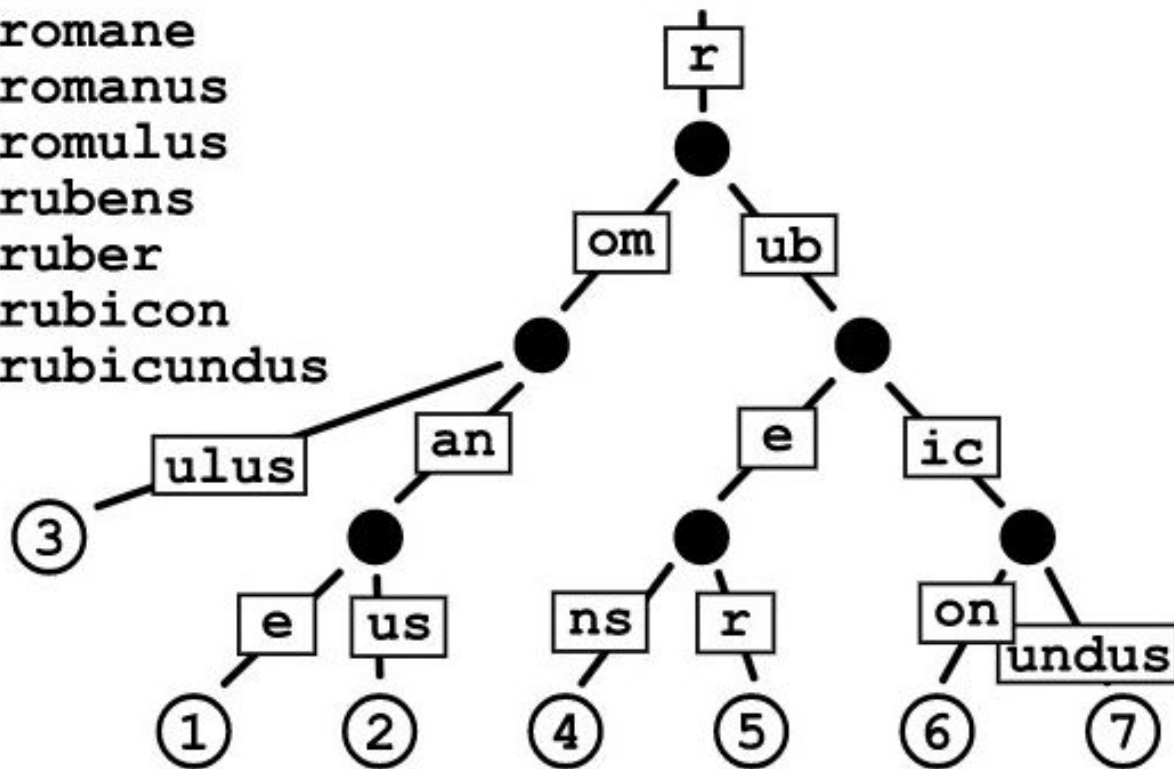
African' iulius	32
A g	
¶ Agamemnon rex grecoꝝ	18
Agar samula sare	5
Agapit' martyr	32
Agapitus martyr	33
Agapes	33
Agapitus papa p̄m'	39
Agapit' papa secundus	51
Ageric' virdunensis ep̄s	40
Aggeus p̄bat	15
Aggeus increpat populū	17
Agiale' p̄m' rex sicionicoꝝ	5
Agilis abbas sc̄us	42
Agilopib' ep̄s colonici.	45
Agatho papa	43
Agatha virgo et martyr	32
Agnes virgo et martyr	33
Agricola martyr	33
Agrippa filius rex italie	11
Agrippina ciuitas	28
A h	
¶ Abab rex isrl̄ idolatra	10
Abalon iudex	8
Hyadan' ep̄s	42
Hymond' rex anglie	49
A l	
¶ Alanus doctor vniuersal'	60
Alba filius rex albanie	9
Alberic' magn'	58
Albertus impator	60
Alterius impator	63
Altert' p̄iar. hierosolimoꝝ	57
Albus' martyr	33
Albus' vir sc̄us	38
Albinus andegauen. ep̄s	45
Alto abbas floziacē.	52
Alchir' ex genē sacerdotali	21
Alcibiades socraticus	17
Alchoran' rex mactomēti	42
Alchuuin' vir coctifim'	46
Alexius vir sanct'	35
Alexander nascif	18
Alexandria ciuitas	18
Alexander magn' regnat	19
Alexader uel Jāne' p̄ntifex	22
Alexander rex syrie	22
Alexandra in linea p̄n.	23
Alexander medic'	30
Alexader magister origenia	30
Alexander mamee impator	31
Alexander ep̄s hierosolimoꝝ	32
martyr	32
Alexander	33
Alexander ep̄s alexādie	34
Alexander de balis doctor	58
Alexander de villa dei	59
Alexander papa p̄m'	25
Alexander papa secund'	54
Alexander papa quart'	62
Alexandria martyr	36
Almanā	31
Almericus h̄retic'	57
Alphonus rex castelle	59
Altinus discipul' sci petri	27
A m	
¶ Amalech delex p̄ saul	9
Amān' ep̄s aureliacē.	36
Amandus vir sc̄us ep̄s	40
Amarias sacerdos	9
Amarias pontifex	12
Amatus exulaf	44
Amazoniū siue massagetarū re-	
gnum oritur	4
Amazones sunt mulieres	8
Amazias rex iuda	12
Ambrosius roman' doctor	35
Amelius comes aluernē.	47
Amic' et berican' martyres	47
Amic' rex assyrioz	7
Aminadab princeps iuda	7
Aminius p̄curator iudee	25
Amichus rex assyrioz	6
Amon filius loth	5
Ammon monachus	36
Amor rex iuda	14
Amos p̄ba	12
Amos pat. 3000. mōachoz	24
Amram filius caath	6
Amr rex isrl̄	10
A n	
¶ Anacle' papa et martyr	28
Ananus rufus	25
Ananias	15
Anastasia virgo	33
Anastasius impator h̄retic'	38
Anastasius impator	45
Anastasius papa p̄m'	36
Anastasi' papa secund'	38
Anastasius papa tert'	56
Anastasi'	50
Anatolia virgo martyr	32
Anatoli' ep̄s antantio.	37
Anaxagoras	17
Anselm' ep̄s cātuarieſ.	54
Anulla	36
Anchises pater enee	8
Anco rex romanoꝝ	15
Andreas ap̄tus	27
Andoen' rotomageſ.	44
Anelinus martyr	31
Anfrid' ep̄s traicetē.	52
Anglia que et britania capis	9
Anglia ad fidē conuertit	30

Общественное достояние,  
<https://commons.wikimedia.org/w/index.php?curid=433142>  
<https://rbscp.lib.rochester.edu/489>



# Term Dictionary

- 1 romane
- 2 romanus
- 3 romulus
- 4 rubens
- 5 ruber
- 6 rubicon
- 7 rubicundus



- rubens
- rub\*
- [rome TO rustic]
- \*uber
- \*man\*

# Offsets to Postings List File

romane	0
romanus	23
romalus	78
rubens	124
rubicon	175
rubicundus	183

10: 8, 9, 10, 14, 18, 23, 24, 26, 31, 35; 8:  
8, 11, 14, 18, 21, 23, 25, 27; 8: 4, 5, 6, 9,  
13, 14, 18, 22; 8: 3, 4, 7, 9, 12, 13, 17,  
20; 7: 5, 9, 12, 14, 19, 23, 28; 9: 0, 2, 5,  
6, 11, 13, 17, 22, 27

# Postings Codecs

□ delta                    8, 9, 10, 14, 18, 23    => 1,1,1,4,4,5

□ vint                    5 => 00000101<sub>2</sub>      129 => 10000001<sub>2</sub> 00000001<sub>2</sub>

□ PFOR                    001<sub>2</sub> 001<sub>2</sub> 001<sub>2</sub> 100<sub>2</sub> 100<sub>2</sub> 101<sub>2</sub>

# Query Execution

romane	0
<b>romanus</b>	<b>23</b>
romalus	78
rubens	124
<b>rubicon</b>	<b>175</b>
rubicundus	183

8, 9, 10, 14, 18, 23, 24, 26, 31, 35

**8, 11, 14, 18, 21, 23, 25, 27**

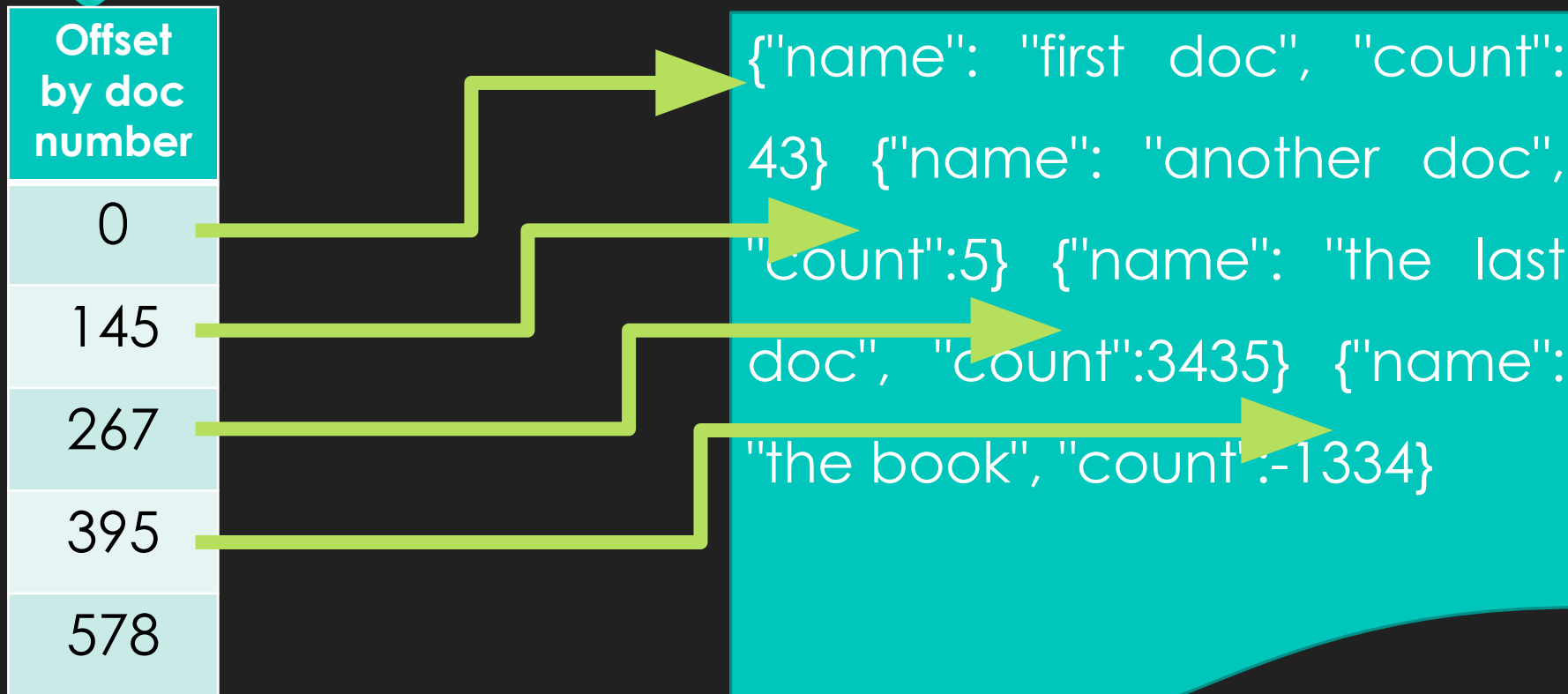
4, 5, 6, 9, 13, 14, 18, 22

3, 4, 7, 9, 12, 13, 17, 20

**5, 9, 12, 14, 19, 23, 28**

0, 2, 5, 6, 11, 13, 17, 22, 27

# Stored Field Retrieval



**Indexing**

# Index Document

PUT ▾

localhost:9200/twitter/tweet/1

Authorization

Headers (1)

Body ●

Pre-request Script

Tests

form-data

x-www-form-urlencoded

raw

binary

JSON (application/json) ▾

```
1 {  
2   "user" : "kimchy",  
3   "post_date" : "2009-11-15T14:12:12",  
4   "message" : "trying out Elasticsearch"  
5 }
```

# Mapping

PUT ▾

localhost:9200/my\_index

Authorization

Headers (1)

Body ●

Pre-request Script

Tests

form-data

x-www-form-urlencoded

raw

binary

JSON (application/json) ▾

```
1 ▾ {
2 ▾   "mappings": {
3 ▾     "properties": {
4       "title":    { "type": "text" },
5       "name":     { "type": "text" },
6       "age":      { "type": "integer" },
7 ▾     "created":  {
8       "type":     "date",
9       "format":  "strict_date_optional_time||epoch_millis"
10      }
    }
  }
```



# Analysis

```
curl -X PUT "localhost:9200/my_index" -H 'Content-Type: application/json' -d'
```

```
{ "settings": {  "analysis": {  
  "analyzer": {  
    "my_custom_analyzer": {      "type":  "custom",  
      "tokenizer": "standard",  
      "filter": [  
        "lowercase"  
      ]  
    }  }  }  }}
```

# In-memory Buffer

```
{  
  "user" : "kimchy",  
  "message" : "trying out  
    Elasticsearch"  
}
```

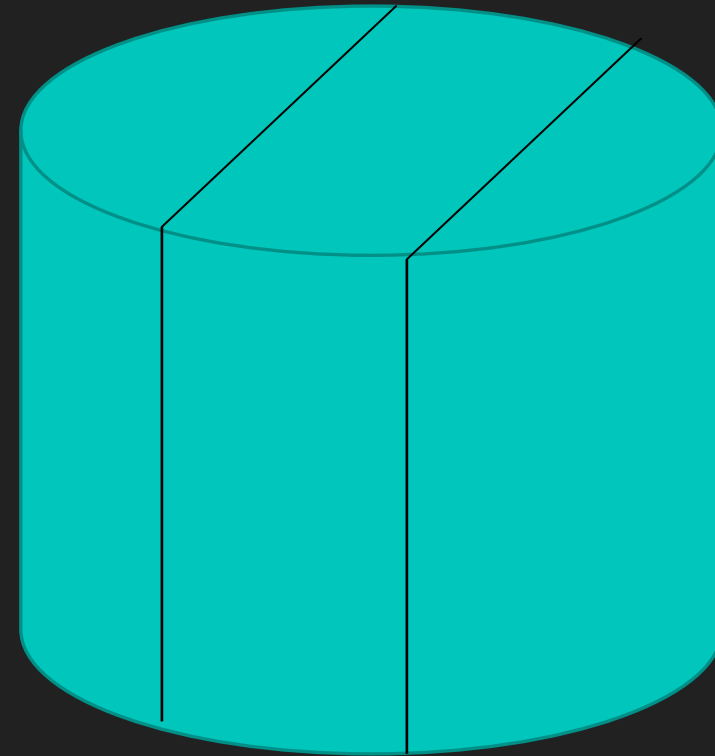
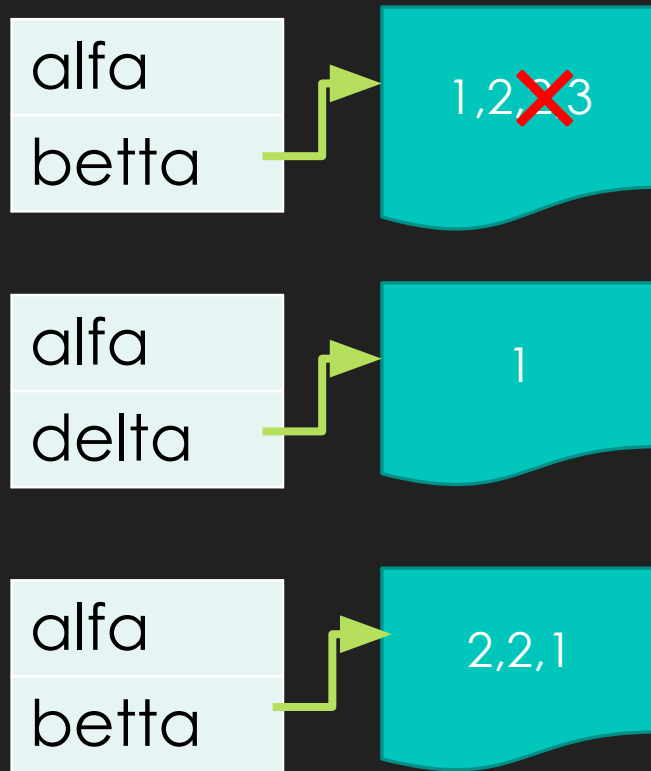
user

kimchy	1	...
--------	---	-----

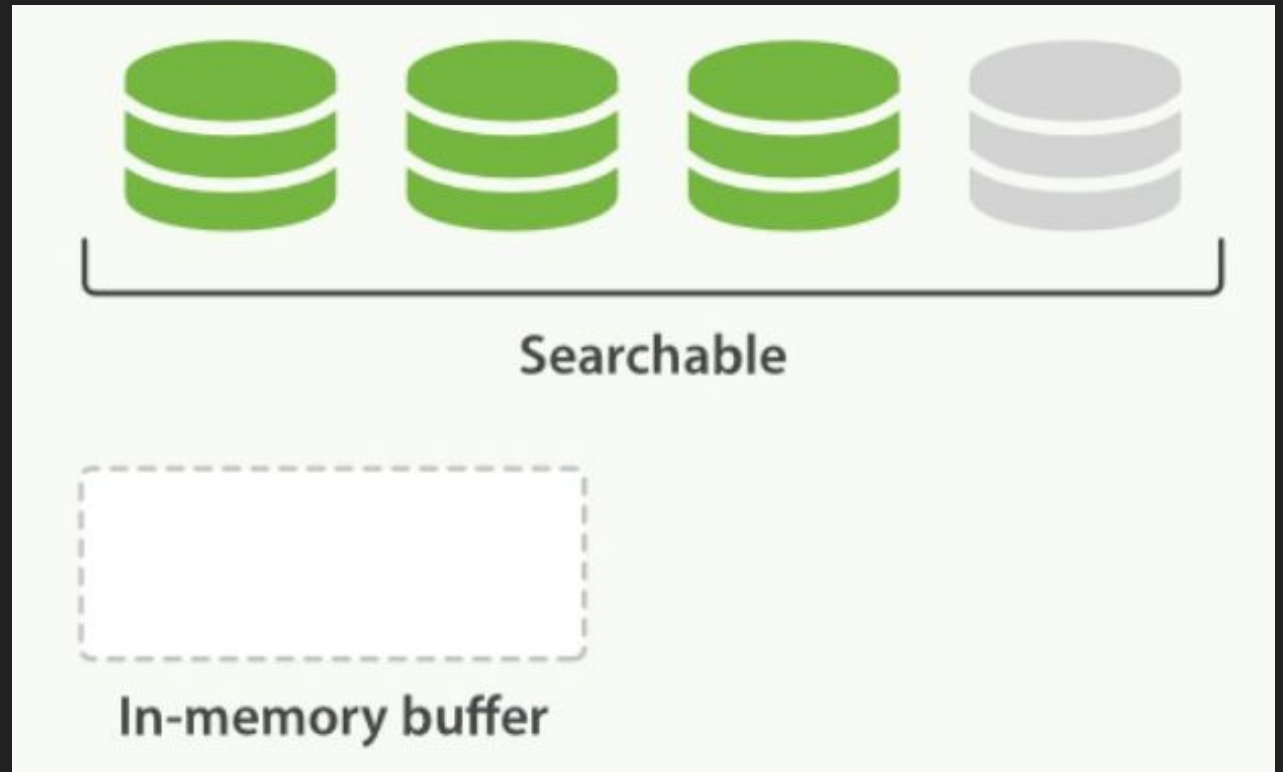
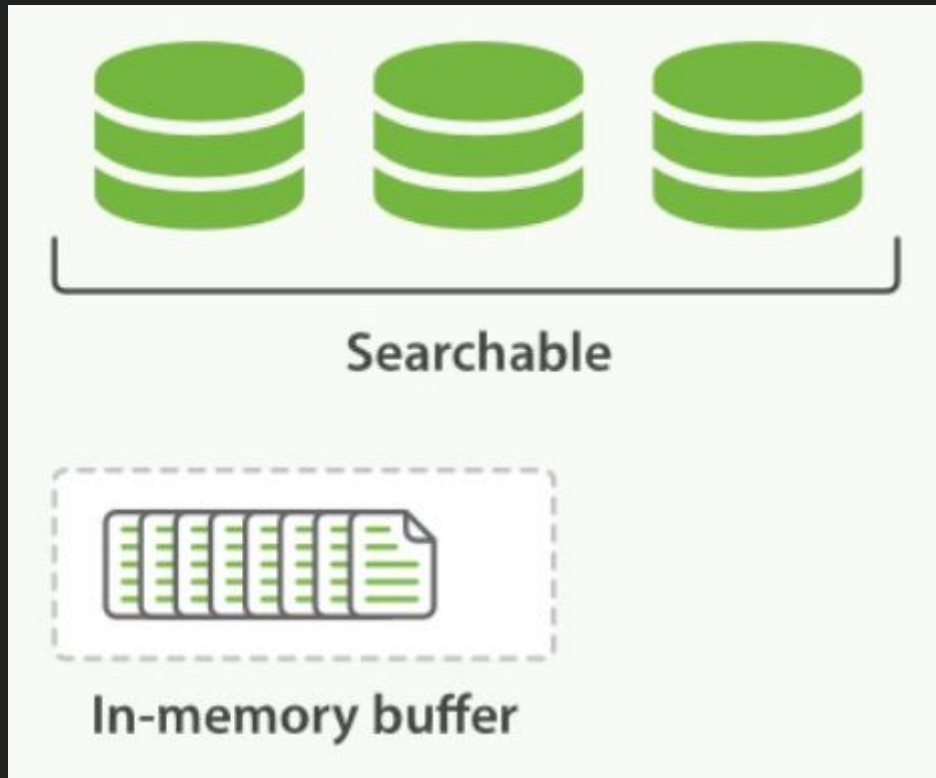
message

elasticsearch	1	...
trying	1	...
out	1	...

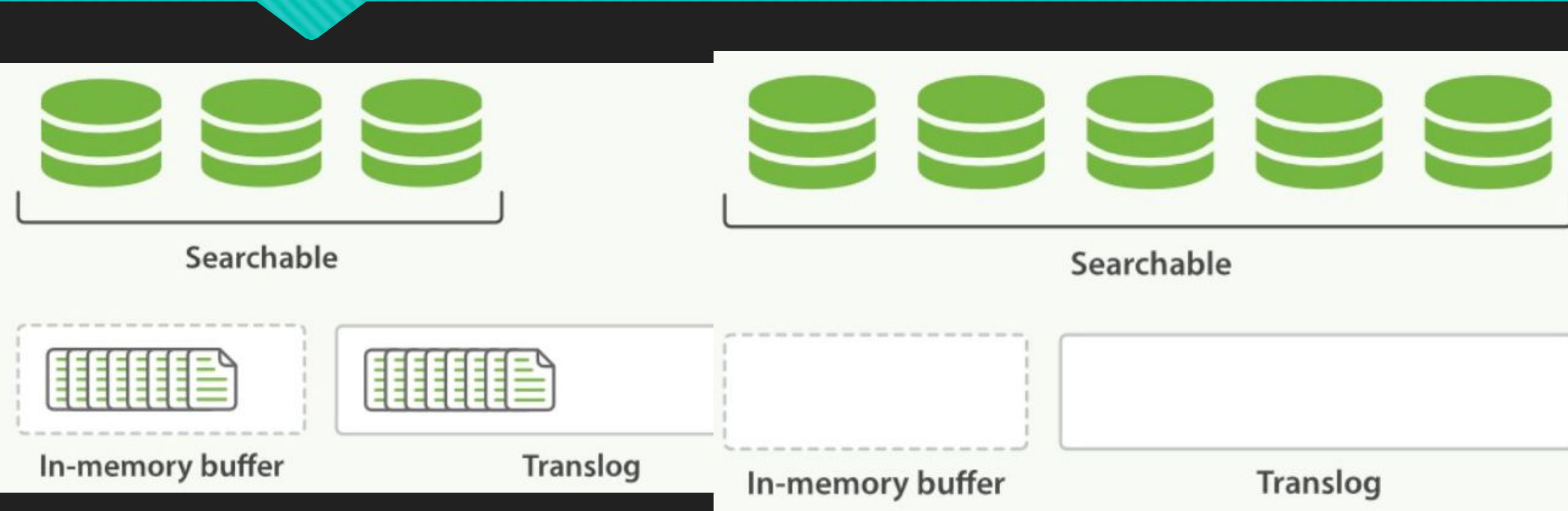
# Segments



# \_refresh

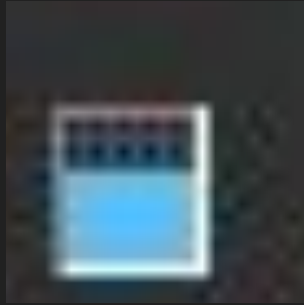


# \_refresh vs \_flush



# Indexing Performance

- `_bulk`
- Threads
- ETL is hard



**Searching**

```
    "skipped": 0,  
    "failed": 0  
  },  
  "hits": {  
    "total": {  
      "value": 23539,  
      "relation": "eq"  
    },  
    "max_score": 1.3862944,  
  }  
}
```

# Result Page

```
"hits" : [  
  {  
    "_index" : "twitter",  
    "_type" : "_doc",  
    "_id" : "0",  
    "_score": 1.3862944,  
    "_source" : {  
      "user" : "kimchy",  
      "date" : "2009-11-15T14:12:12",  
      "message" : "trying out Elasticsearch",  
      "likes": 0  
    }  
  }  
]
```



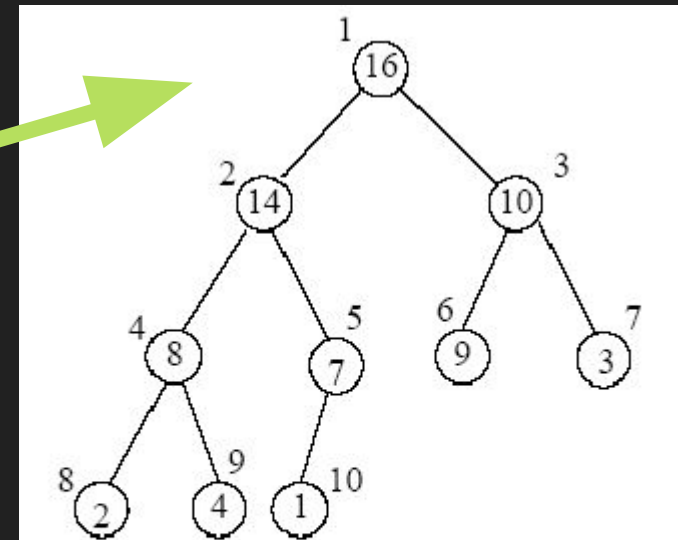
# Result Page Cropping

10: 8, 9, 10, 14, 18, 23, 24, 26,

31, 35; 8: 8, 11, 14, 18, 21, 23,

25, 27; 8: 4, 5, 6, 9, 13, 14, 18,

22; 8: 3, 4, 7, 9, 12, 13, 17, 20;



$O(n \log(p))$

[http://www.cse.hut.fi/en/research/SVG/TRAKLA2/tutorials/heap\\_tutorial/taulukkona.html](http://www.cse.hut.fi/en/research/SVG/TRAKLA2/tutorials/heap_tutorial/taulukkona.html)

1	2	3	4	5	6	7	8	9	10
16	14	10	8	7	9	3	2	4	1

# Мобильные телефоны

Сортировать: по цене

по цене

Price by  
Doc  
Number

по размеру скидки по новизне

Сначала предложения в моём



Телефон Digma LINX A105N 2G

390 ₺

390

4466

12453

390

2342

390

390

2341



Телефон Ji

390 ₺



Телефон Digma LINX A105 2G

396 ₺

## Size

## Heel Height

- Low 1-2" (120)
- Mid 2-3" (399)
- High 3-4" (415)
- Ultra High 4" & Over (157)

## Heel Type

- Block (306)
- Cone (16)
- Kitten (117)
- Sculpted (10)
- Stacked (152)
- Stiletto (352)
- Wedge (42)

## Brand

🔍 Search by brand

- Adrianna Papell (17)
- Adrienne Vittadini (8)
- Aerosoles (14)

```
{
  "aggs" : {
    "genres" : {
      "terms" : { "field" : "genre" } ❶
    }
  },
  ...
  "aggregations" : {
    "genres" : {
      "doc_count_error_upper_bound": 0, ❷
      "sum_other_doc_count": 0, ❸
      "buckets" : [ ❹
        {
          "key" : "electronic",
          "doc_count" : 6
        },
        {
          "key" : "rock",
          "doc_count" : 3
        },
        {
          "key" : "jazz",
          "doc_count" : 2
        }
      ]
    }
  }
}
```

<https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-terms-aggregation.html>

**Real Life Usage**

type: http AND http.code: 302

Uses lucene query syntax



Add a filter +

packetbeat-\*

Selected Fields

? \_source

Available Fields

@timestamp

t \_id

t \_index

# \_score

t \_type

# bytes\_out

t client\_ip

client\_location

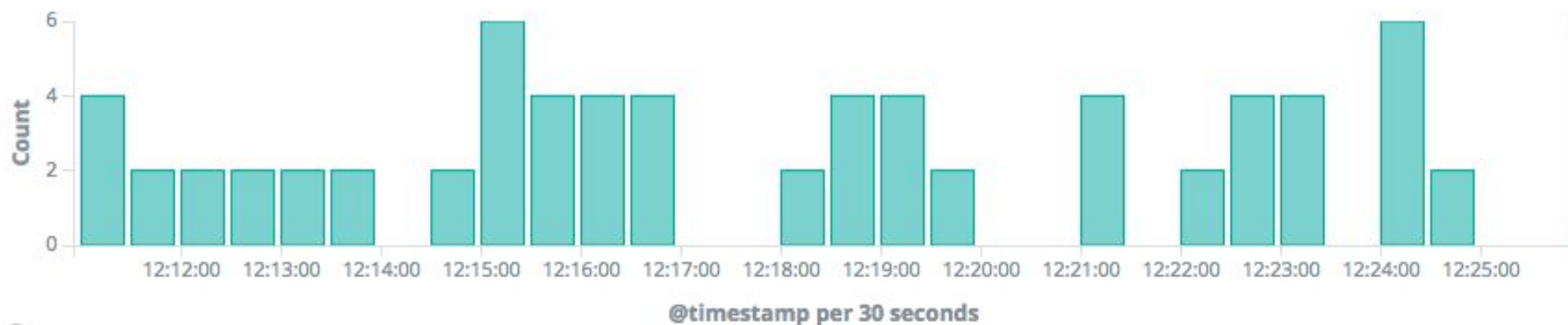
# client\_port

t client\_server

# count

March 14th 2018, 12:10:55.313 - March 14th 2018, 12:25:55.313

Auto



Time

\_source

- ▶ March 14th 2018, 12:24:59.000
- ```

type: http status: Error client_ip: 127.0.0.1 http.content_length: 209 http
.phrase: FOUND http.code: 302 http.response_headers.content_type: text/html; ch
arset=utf-8 http.request_headers.host: packetbeat.com timestamp: February 10th 2
014, 11:27:38.276 @timestamp: March 14th 2018, 12:24:59.000 client_port: 56,341
query: GET /logout HTTP/1.1 path: /logout server: app.server4 response: HTTP/

```
- ▶ March 14th 2018, 12:24:59.000
- ```

type: http status: Error client_ip: 186.216.161.91 http.content_length: 209
http.phrase: FOUND http.code: 302 http.response_headers.content_type: text/html
; charset=utf-8 http.request_headers.host: packetbeat.com timestamp: February 10
th 2014, 11:27:38.276 @timestamp: March 14th 2018, 12:24:59.000 client_port: 56,
341 query: GET /logout HTTP/1.1 path: /logout server: nginx-proxy2 response:

```

# Scaling

## Original Table

CUSTOMER ID	FIRST NAME	LAST NAME	FAVORITE COLOR
1	TAEKO	OHNUKI	BLUE
2	O.V.	WRIGHT	GREEN
3	SELDA	BAĞCAN	PURPLE
4	JIM	PEPPER	AUBERGINE

## Vertical Partitions

VP1

CUSTOMER ID	FIRST NAME	LAST NAME
1	TAEKO	OHNUKI
2	O.V.	WRIGHT
3	SELDA	BAĞCAN
4	JIM	PEPPER

VP2

CUSTOMER ID	FAVORITE COLOR
1	BLUE
2	GREEN
3	PURPLE
4	AUBERGINE

## Horizontal Partitions

HP1

CUSTOMER ID	FIRST NAME	LAST NAME	FAVORITE COLOR
1	TAEKO	OHNUKI	BLUE
2	O.V.	WRIGHT	GREEN

HP2

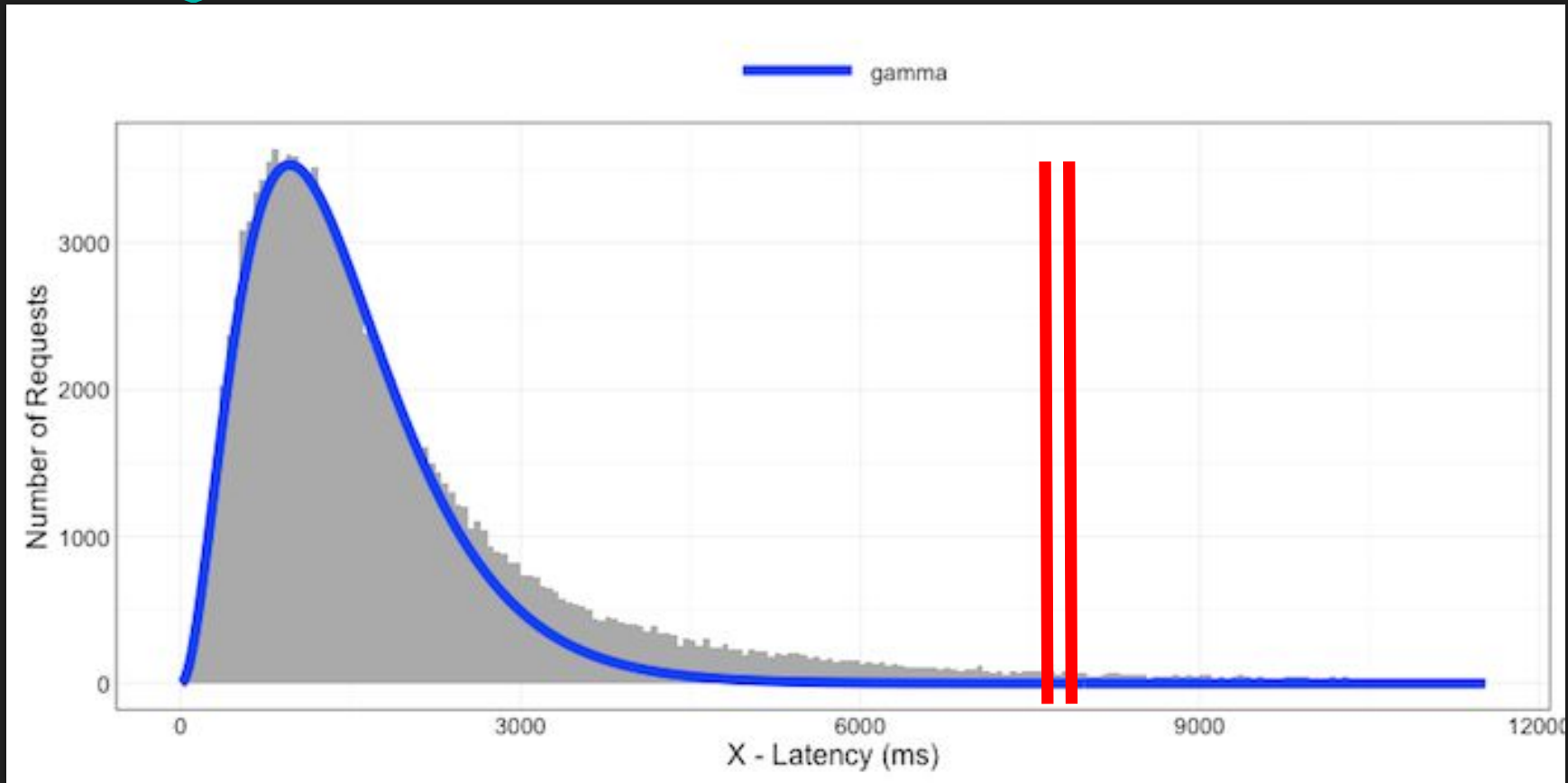
CUSTOMER ID	FIRST NAME	LAST NAME	FAVORITE COLOR
3	SELDA	BAĞCAN	PURPLE
4	JIM	PEPPER	AUBERGINE

<https://www.digitalocean.com/community/tutorials/understanding-data-base-sharding>

# Summary

- Why search?
- Inverted Index
- Indexing
- Searching
- Scaling
- ELK

# 200 threads





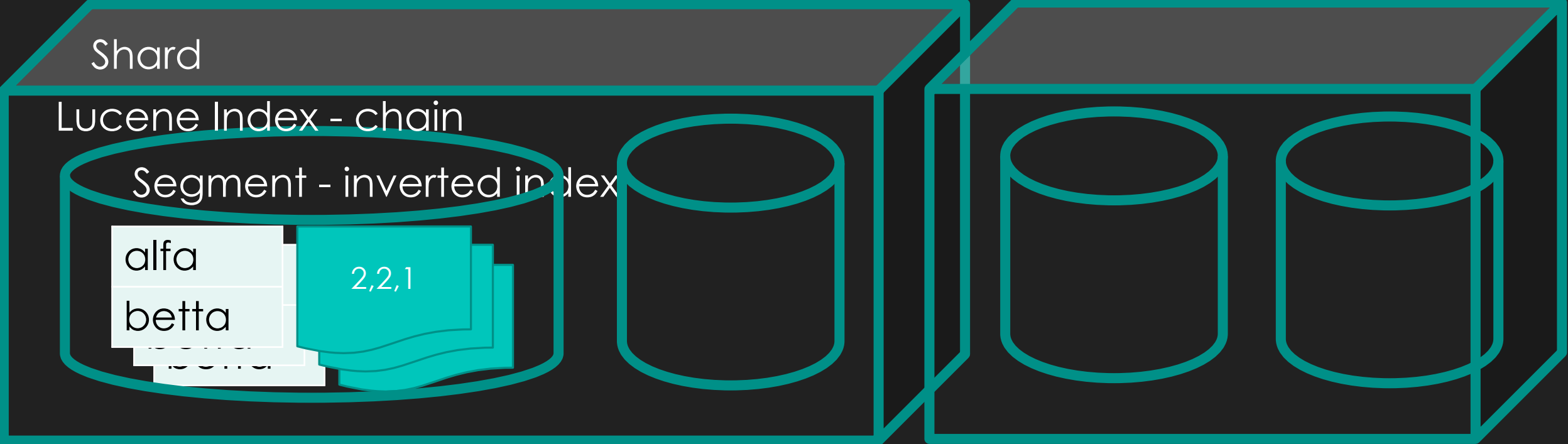
# Index Hierarchy

Index pattern: access-log-\*

- access-log-2019-08-06
- access-log-2019-08-07
- access-log-2019-08-08

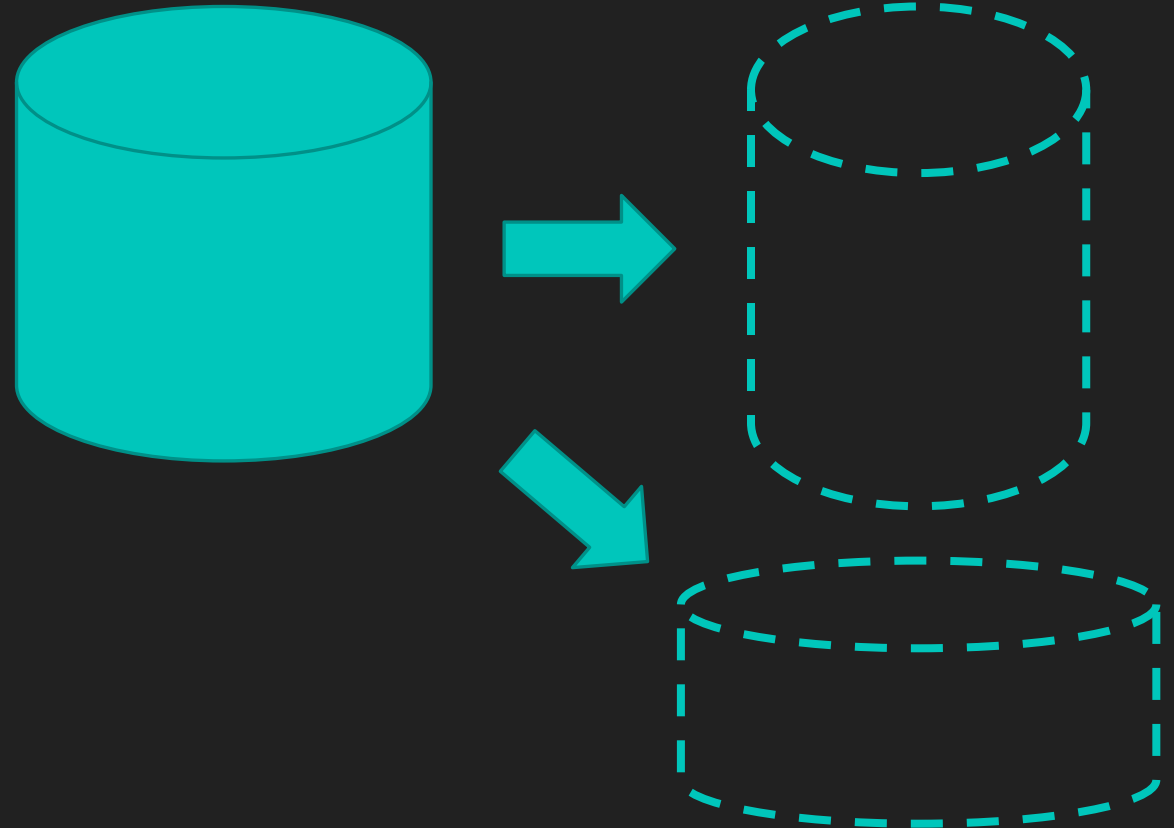
....

Elasticsearch Index



# An Index is ...

- an derived data structure



romane

**romanus**

romalus

rubens

**rubicon**

rubicundus

8, 9, 10, 14, 18

**8, 11, 14, 18, 21**

4, 5, 6, 9, 13, 14

3, 4, 7, 9, 12, 13

**5, 9, 12, 14, 19**

0, 2, 5, 6, 11, 13

# Term Frequency

red bar

red red  
red bar  
red

# Inverse Document Frequency

Chicago  
Bulls

Red  
Socks

**Computing is too important to be left to men**

Karen Spärck-Jones

