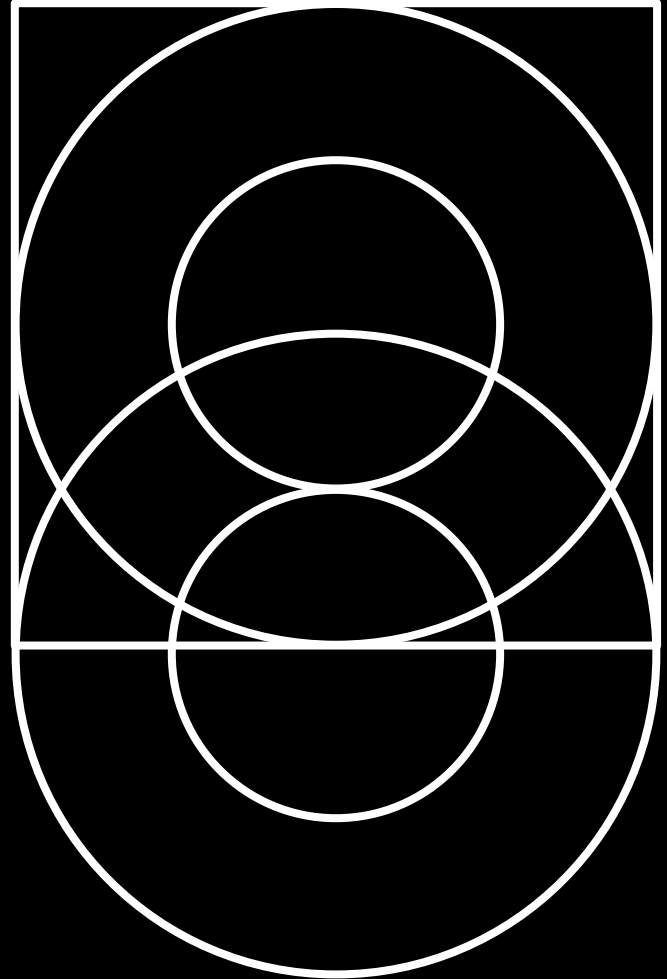




Building a domain specific search engine

digital natives

unit8.co



■ **Great search engines are everywhere**
(since quite some time now ...)

Google



Google Search

I'm Feeling Lucky

1-24 of over 50,000 results for **Cell Phones & Accessories : Cases, Holsters & Clips** : "iphone case"

Sort by Suggested ▾

☐ **FREE Shipping**

All customers get **FREE Shipping** on orders over \$25 shipped by Amazon

Show results for

Any Category

Cell Phones & Accessories

Cases, Holsters & Clips

Cases

Wallet Cases

Flip Cases

Waterproof Cases

⌵ [See more](#)

Refine by

Amazon Prime☐ **prime**

SPONSORED BY LIFEPROOF

**Water, drop, dirt and snow
proof for iPhone 7**Shop now ›Lifeproof FRĒ SERIES Waterpro...
✓prime ★★★★★ 1,561Lifeproof FRĒ SERIES Waterpro...
✓prime ★★★★★ 1,561[Ad feedback](#)Showing results in **Cell Phones & Accessories**. Show instead results in [All Departments](#).

UBER 10:09 AM 100%



301 Front St W



Search for a restaurant or cuisine

— RESTAURANTS AS SOON AS 15 MIN —



Electric Mud

15-26 MIN

Canadian / American • Best of Toronto • \$\$



Gusto 101

19-30 MIN

Italian • Best of Toronto • \$\$



Q Try "Amsterdam"

EXPLORE AIRBNB

All

Stays

Experiences

Adventures

Restaurants

RECENT SEARCHES

- 📍 Espoo · Logements
Nov 15 - 18
- 📍 Espoo
Nov 15 - 18
- 📍 Valais · Logements
Dec 28, 2019 - Jan 02, 2020 · 10 guests
- 📍 Édimbourg · Logements
Dec 28, 2019 - Jan 02, 2020
- 📍 Édimbourg
Dec 28, 2019 - Jan 02, 2020

Become a host

Saved

Trips

Messages

Help



Adventures



Restaurants



Edinburgh, United Kingdom

Dec 28 - Jan 02

 **Yet, many enterprise tools
seem to neglect it**

[Task Lists](#) [Landscape Overview](#) [Critical Objects](#) [Cross-System Object Locks](#) [Transport Analysis](#)**Search**[All Content](#) [IT Service Management](#) [Solution Documentation](#)**Advanced Search**Search For: ***dolores***Search In: **Change Request Management: Transaction** ▾**Search**[Basic Search](#)**Additional Search Criteria**

Object ID	▾	Equal To	▾	<input type="text"/>			
Transaction Type	▾	Equal To	▾	<input type="text"/>			
Description	▾	Contains All of These Words	▾	<input type="text"/>			
Created on	▾	Equal To	▾	<input type="text"/>			

[Home](#) > [Change Request Management: Transaction](#)[All Content](#)[Change Request Management...](#)Sort By: **Relevance** ▾

	Dolores sol, 8000000263	Change Request Management: Transaction
	Transaction Type Description: Phase Cycle	Priority: 4: Low
	Current Processor:	Last Changed by:
	Configuration Item:	Change Cycle: 8000000263
	Close Details	
	test uc nc, 8000000479	Change Request Management: Transaction
	Transaction Type Description: Request for Change	Priority: 4: Low
	Current Processor:	Last Changed by:
	Configuration Item:	Change Cycle: 8000000263
	Details	
	test ES, 8000000440	Change Request Management: Transaction

[Close](#)Details For **Dolores sol, 8000000263**

Due by:

Transaction GUID: **005056B510841**

Transaction Type: **SMIM**

System Status: **Open**

Created on: **18.04.2016**

Last Changed at: **18.04.2016**

Impact:

Urgency:

Requested Start: **18.04.2016**

Requested End: **21.04.2016**



Activities

SuiteView

Quick Start

Transactions

Lists

Reports

Customization

Documents

Setup

Support

Sales

Knowledge Base

Inventory Item

List Search Customize More

Save



Cancel

Reset

Primary Information

FORM

Primary Inventory Part Form

SUBITEM OF

<Type then tab>

ITEM NAME/NUMBER *

9632

UPC CODE

452165

DISPLAY NAME/CODE

Standard Keyboard

UNITS TYPE

|

Area

Box

Ea/Bx (12)/Cs(48)

Ea/Case(12)

Each

Length

ITEM IMAGE

<Type then tab>

Inventory

Pricing

Accounting

System Information

Store

Warranty Information

SuiteOffice - Badges

SuiteOffice - Image Gallery



Item/Cost Detail

COSTING METHOD

Average

TRANSFER SHIP PRICE

COST CATEGORY *

Default Cost Category

PURCHASE PRICE

☐ AVAILABLE TO PARTNERS☐ OFFER SUPPORT☐ TRACK LANDED COST☐ DROP SHIP ITEM



compliance antan



demo_legal_matter

Welcome,
Yves



Repository File server SharePoint Websites Only matters

100%

Antan Industries Carol Smith

08-01-2018 21:01:16

709 2018-01-08 Antan Industries Antan Industries Consectetuer Avenue 96 2 6915... Tintigny DE 08 09 95 45 42 Carol Smith F EN Compliance Officer 08 09 95 45 42 carol... Antan Industries Carol Smith

17.1%

Deloitte Academy Webinar - Pan-European VAT Update - Q2 _ Q3 2015 - 14_09_2015_1...

09-01-2018 15:19:35

to ensure compliance with national legislations. Agenda * VAT changes

15.6%

Antan Industries - Trade Agreement - 20170906 - Version 3.docx

09-01-2018 15:50:23

Director, (hereafter "Knowliah") and, Antan Industries with registered... acting in his or her capacity as Compliance Officer, (hereafter "Client"...-User will refer to that legal entity. The Client may use the Services only in compliance... Cloud System in relation to the Client Materials; compliance of the Client... in accordance with the instructions of the Client; if it cannot provide such compliance... Antan Industries - Trade Agreement - 20170906 - Version 3.docx

15.6%

Antan Industries - Trade Agreement - 20170906 - Version 3.docx

09-01-2018 22:40:29

Director, (hereafter "Knowliah") and, Antan Industries with registered... acting in his or her capacity as Compliance Officer, (hereafter "Client"...-User will refer to that legal entity. The Client may use the Services only in compliance... Cloud System in relation to the Client Materials; compliance of the Client... in accordance with the instructions of the Client; if it cannot provide such compliance... Antan Industries - Trade Agreement - 20170906 - Version 3.docx

13.4%

Antan Industries - Trade Agreement - 20170906 - Signed.pdf

09-01-2018 15:50:27

, Antan Industries with registered offices at Consectetuer Avenue, 6915 Tintigny... 75709925399 represented by Carol Smith acting in his or her capacity as Compliance... use the Services only in compliance with these Terms & Conditions. The Client may... Materials; 6.1.2. compliance of the Client Materials with applicable laws... with the instructions of the Client; if it cannot provide such compliance... Antan Industries - Trade Agreement - 20170906 - Signed.pdf

1

Page 1 of 1 (5 results)
There are in total 10 result(s).

extension [4]

docx [2]

contact [1]

msg [1]

pdf [1]

Information_Type [3]

contract [3]

contact datasheet [1]

correspondence [1]

Period [5]

Past 24 hours

Past week

Past month

Past year

Older...

Why enterprises should care?

- The average company is using >1000 cloud services which can be hard to unify and search the relevant information
- Many workers rely on quickly finding the relevant pieces of information (Customer support, Sales, Ops, ...)
- It has become easy to deploy and customize existing open-source search engines



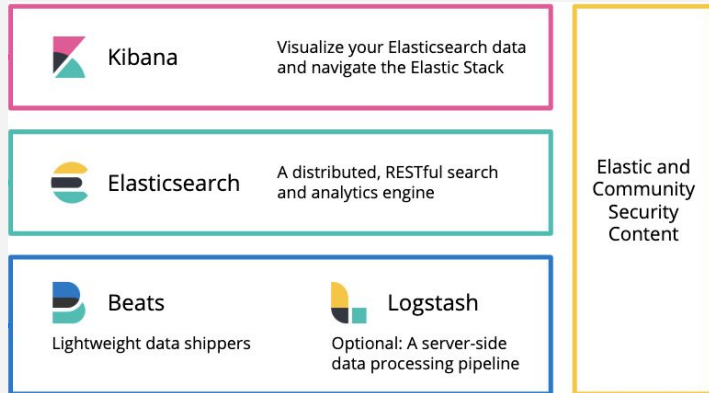


Building your own search engine

(using Elasticsearch)

10sec. intro to Elasticsearch

- Document based
- Search and analytics engine
- Distributed and scalable
- Easy to ingest many sources




Terminology in Elasticsearch

Table → Index

Row → Document

Column → Field

Schema → Mapping



Understand your search engine

(& how to get the most out of it)


Major parts of a search engine

1. Split sentences into words (**Tokenizer**)
2. Index words in documents (**Indexing**)
3. Rank documents (**Ranking**)



Let's build ours!

(<https://github.com/unit8co/dmz-workshop-er-search>)

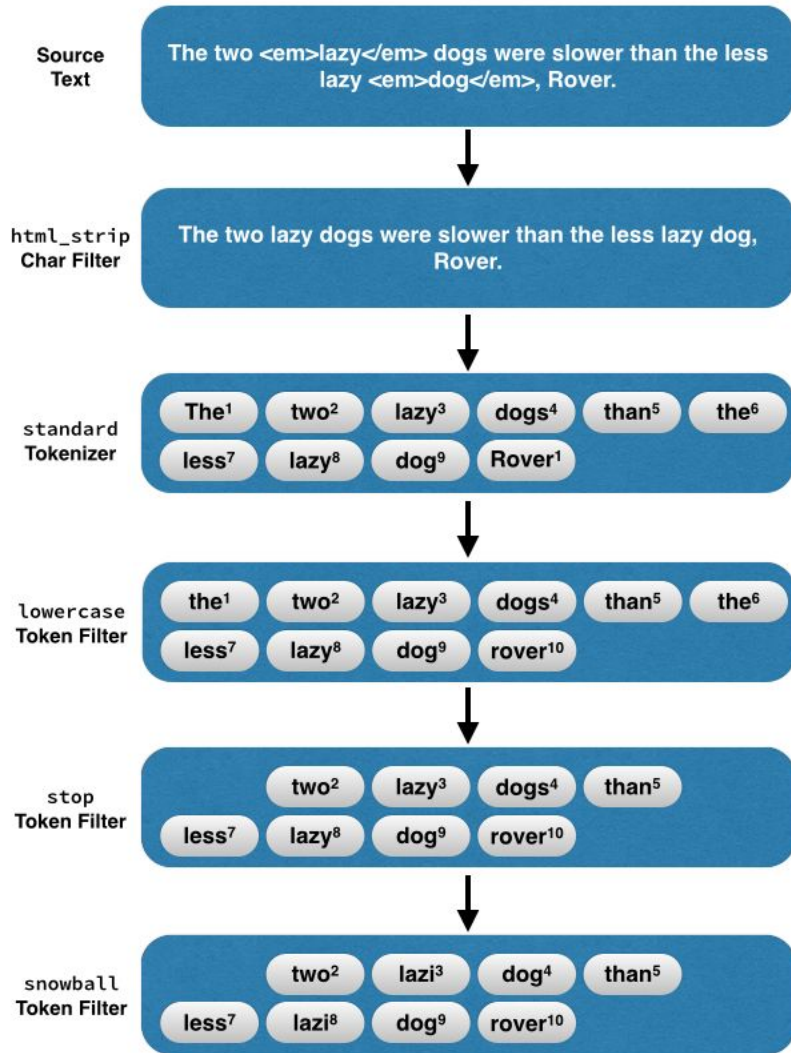


Tips to customize your **Tokenizer**

(or should we say analyzer)

Tips #1: Analyzers are powerful

- It is worth tailoring the analyzer for your most important fields:
 - Language (e.g. “Balletttänzerin”)
 - Field content (e.g. parsing IDs)
 - Case sensitive?
 -



Tips #2: Synonyms

- Great tool to reconcile different notations across documents:
 - Leverage existing business knowledge and abbreviations
 - Use existing ontologies

```
{  
  "region_synonym": {  
    "type": "synonym",  
    "synonyms": [  
      "europe, eu, eur",  
      "middle-east, me, moyen-orient",  
      "india middle east africa, ima",  
      "latin america, south america, lam, latam",  
      "north america, nam",  
      "south east asia, sea",  
      "north east asia, nea",  
      "united states of america, usa, us",  
      "united kingdom, uk",  
      "all regions, world, monde"  
    ]  
  }  
}
```

A few examples of real-life issues

- “-” (dash) vs “—” (em-dash)
 - **Possible solution:** normalize characters in analyzers
- Special Characters in IDs
 - DEF#265/312
 - ABC*234/124
 - **Possible solution:** Change the analyzer for that field
- Language specific issues
 - “L'Équipe” vs “L Equipe”
 - “Balletttänzerin”
 - **Possible solution:** Pick the proper language analyzer
 - For German, have a look at “Word decomposers”

What we don't have time to talk about

- Shingle (index groups of words for faster queries)
- Phonetic analyzers
- Pattern Capture
- ...

If you want to know more, please have a look at:

<https://www.elastic.co/blog/found-text-analysis-part-1>

Tips to customize your Ranking

Tips #1: Understanding ranking is key

$$\text{score}(q,d) = \text{coord}(q,d) \cdot \text{queryNorm}(q) \cdot \sum_{t \text{ in } q} (\text{tf}(t \text{ in } d) \cdot \text{idf}(t)^2 \cdot t.\text{getBoost}() \cdot \text{norm}(t,d))$$

tf = Measure of how often a [term appears in the document](#)

idf = Measure of how often the [term appears in all documents](#)

norm = Normalization factor based on number of words in field

boost = Boost of the field

coord = Number of [terms \(in the query\) found in the document](#)

queryNorm = Normalization query score (so it can be compared across queries)

What this means in practice

- Documents containing all the search terms are good
- Matches on rare words are better than for common words
- Long documents are not as good as short ones
- Documents which mention the search terms many times are good

Tips #2 Match the full sentence

- Without quotes exact matches do not have a special importance
 - Use a query to match the words (**match**) and the the full sentence (**match_phrase**) at the same time

event sourcing using kafka

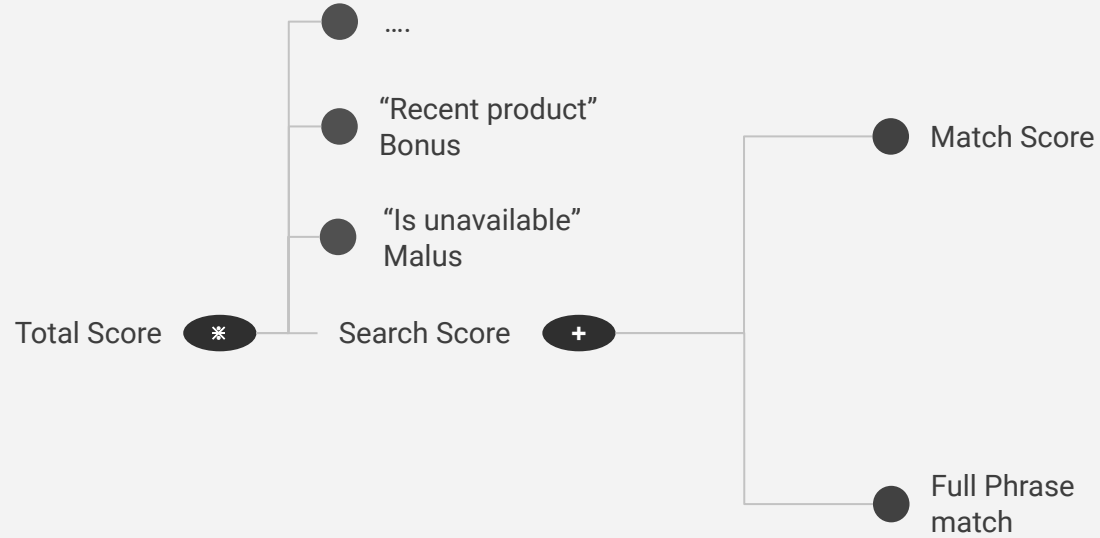


1. **Kafka** — how to configure
2. New release of Apache **Kafka**
3. **Event Sourcing using Kafka**
4. **Event** Store — lessons learnt

Tips#3: Function score

- Use the function score which allows to perform extra computation on document score:
 - Boost more recent products (e.g. last Mercedes S)
 - Promote “popular” items based on your fancy machine learning model
 - ...

```
GET /_search
{
  "query": {
    "function_score": {
      "query": {
        "match": {
          "message": "elasticsearch"
        }
      },
      "script_score" : {
        "script" : {
          "Source":
            "Math.log(2+doc['likes'].value)"
        }
      }
    }
  }
}
```



 **Lots of parameters to tune.**
How do we evaluate our changes?

Ranking Evaluation API

- Allow to give score to specific search results
 - Great way to agree with business on important “top results”
- Great way to validate changes to analyzers/ranking
- Can integrate user feedback directly there

<https://www.elastic.co/blog/made-to-measure-how-to-use-the-ranking-evaluation-api-in-elasticsearch>

POST `/enwiki_rank/_rank_eval`

...

```
"query": {  
  "query_string": { "query": "JFK" }  
}
```

...

```
"ratings": [  
  { "_id": "3054546", "rating": 1 },  
  { "_id": "5119376", "rating": 3 },  
  ...  
]
```

...

```
"metric": { ... }
```



Manage access to data

X-Pack Security features

To operate your elasticsearch cluster, X-Pack has a lot of easy-to-use security features for free (since May 2019):

- TLS encryption
- Role Based access control
- Key management



What's next ?

(In the world of search engines)

Entity Based Search

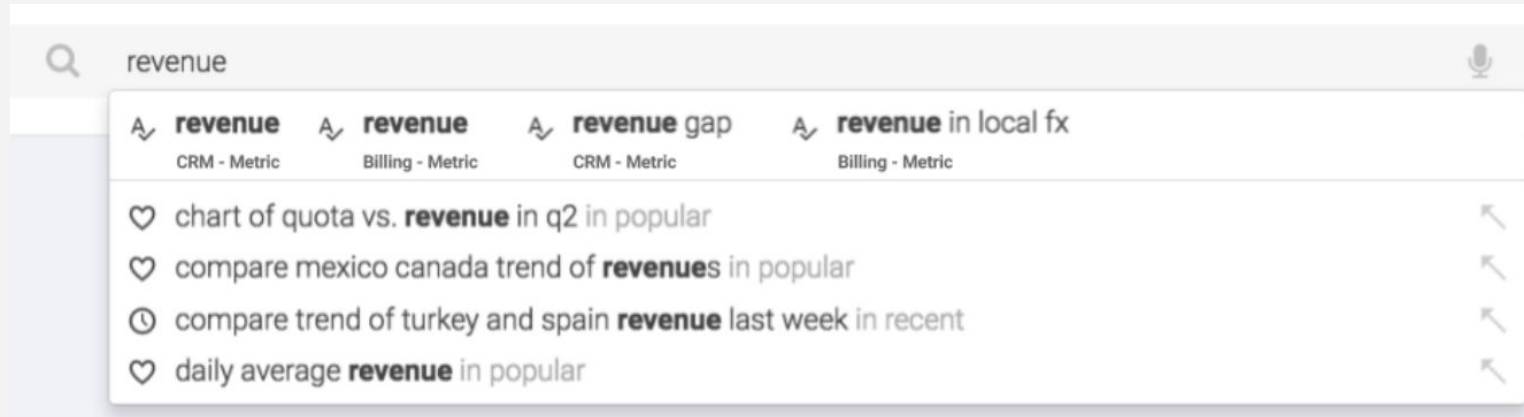
- Improve search by disambiguating some user keywords
- Highlight words belonging to an ontology

<https://www.elastic.co/blog/search-for-things-not-strings-with-the-annotated-text-plugin>

POST my_index/_analyze

```
{  
  "field": "my_rich_text_field",  
  "text": "Today [Johannes](Johannes Mueller)  
          announced a new model of  
          electrical car"  
}
```

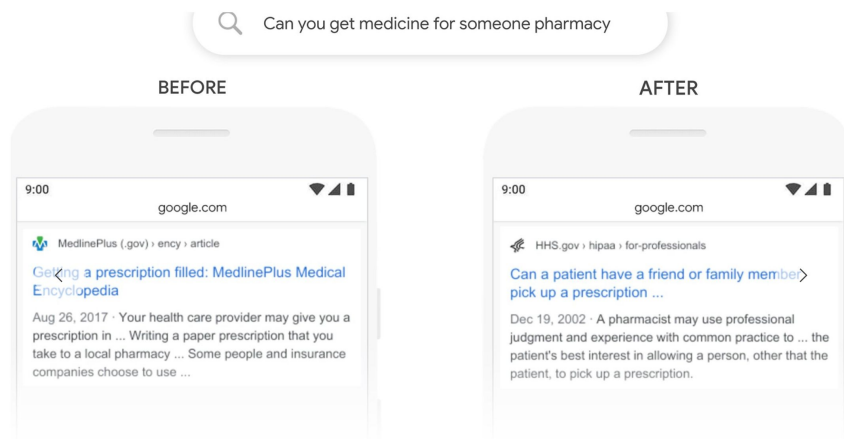
Entity Based Search



Understand user intent

- Integrate neural networks embedding (BERT) in ranking to return better results
- Easy to transfer learning to other languages than English

<https://blog.google/products/search/search-language-understanding-bert/>

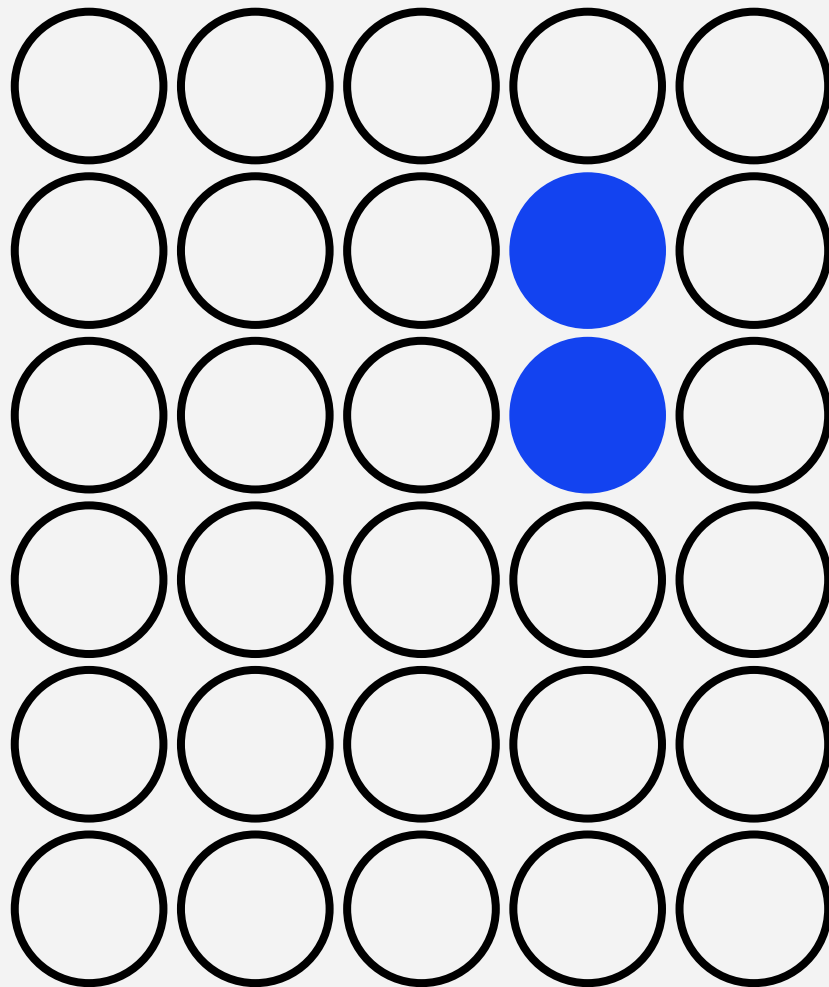


Conclusions

- Running (e.g. managed instance on the cloud) and configuring Elasticsearch
- Including your expertise or knowledge can help providing relevant results
- With the progress in NLP, probably lots of improvements coming in the next few years!

unit8.co

**thank
you**



Useful Links

- Elasticsearch documentation:
<https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html>
- Elasticsearch intro to analyzers:
<https://www.elastic.co/blog/found-text-analysis-part-1>

Steps to search engines

