jobrapido

Elastic Stack in A Day
Milano – 16 Giugno 2016

REVOLUTIONIZE THE WAY PEOPLE GET JOBS WITH ELASTICSEARCH

elastic

SEACOM
the leading open source architects

ABOUT ME

NAME
Salvatore Vadacca

ROLE
Head of Technology @ Jobrapido

EMAIL
salvatore.vadacca@jobrapido.com

TWITTER
@totovadacca

LINKEDIN
https://it.linkedin.com/in/salvatorevadacca

COMPANY WEBSITE
www.jobrapido.com

jobrapido

WHO WE ARE

Jobrapido is the world's leading jobsearch engine that analyses and collects all job posts on the web, giving jobseekers all offers available, ordered for relevance based on the search they've done

Jobrapido

Response

Aggregation

Analysis

VISITORS
1.0 BN visits / year

UNIQUE VISITORS
35 Mio Uvs / month

SUBSCRIBERS
60+ Mio subs users (current stock)

PAGEVIEWS / CLICKS*
280 Mio PVs / month & 130 Mio clicks / month

JOBS
20+ Mio jobs at any given time

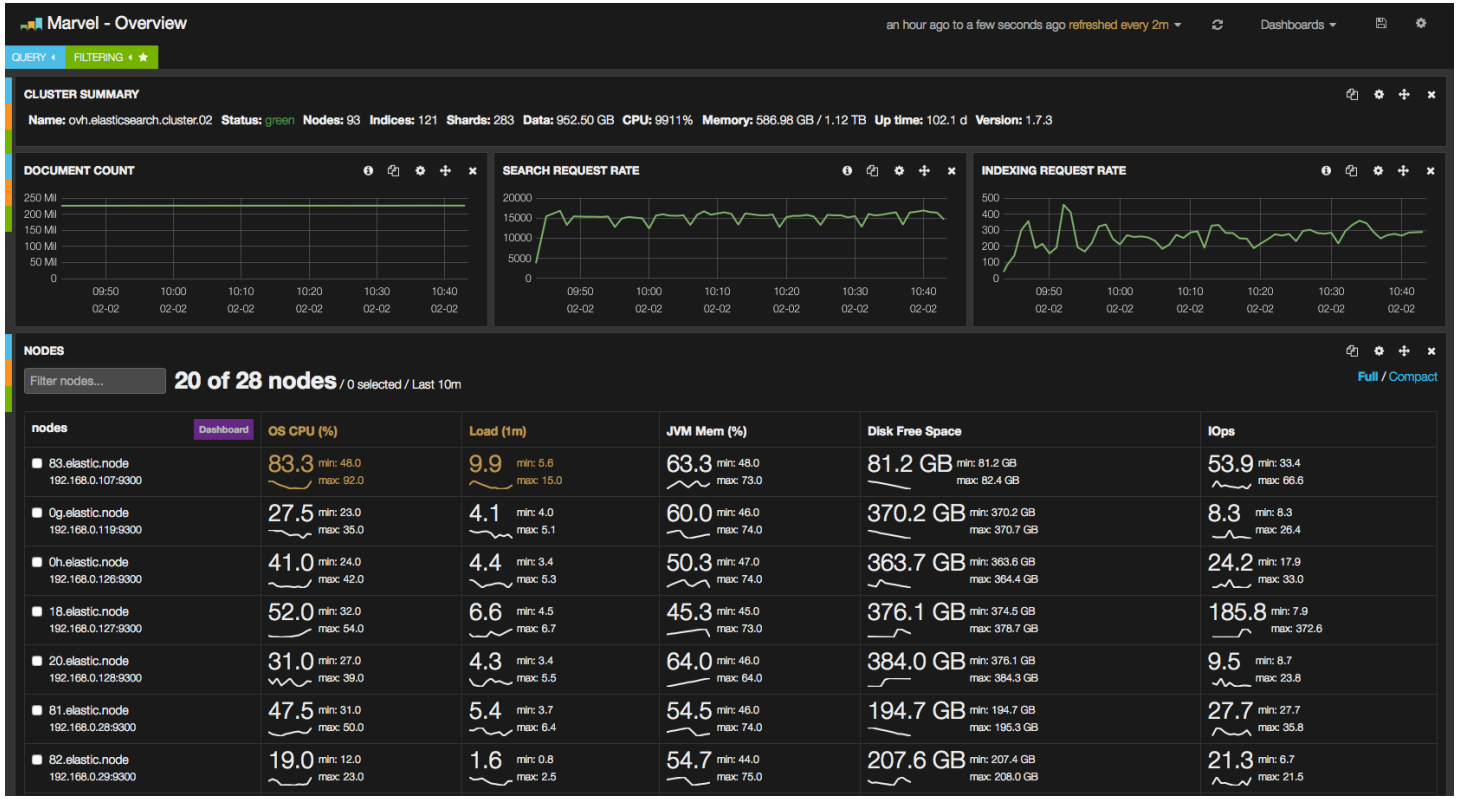
WEBSITES IN 58 COUNTRIES
Head office Milan + office in Amsterdam

PEOPLE
100+

SEACOM
the leading open source architects

jobrapido

OUR ELASTIC NUMBERS (MAIN CLUSTER)



NODES
25 data, 3 masters, 65 client

DATA
1TB

AVG SEARCH RATE
20K/sec

AVG INDEX RATE
400/sec

MEMORY
>1TB

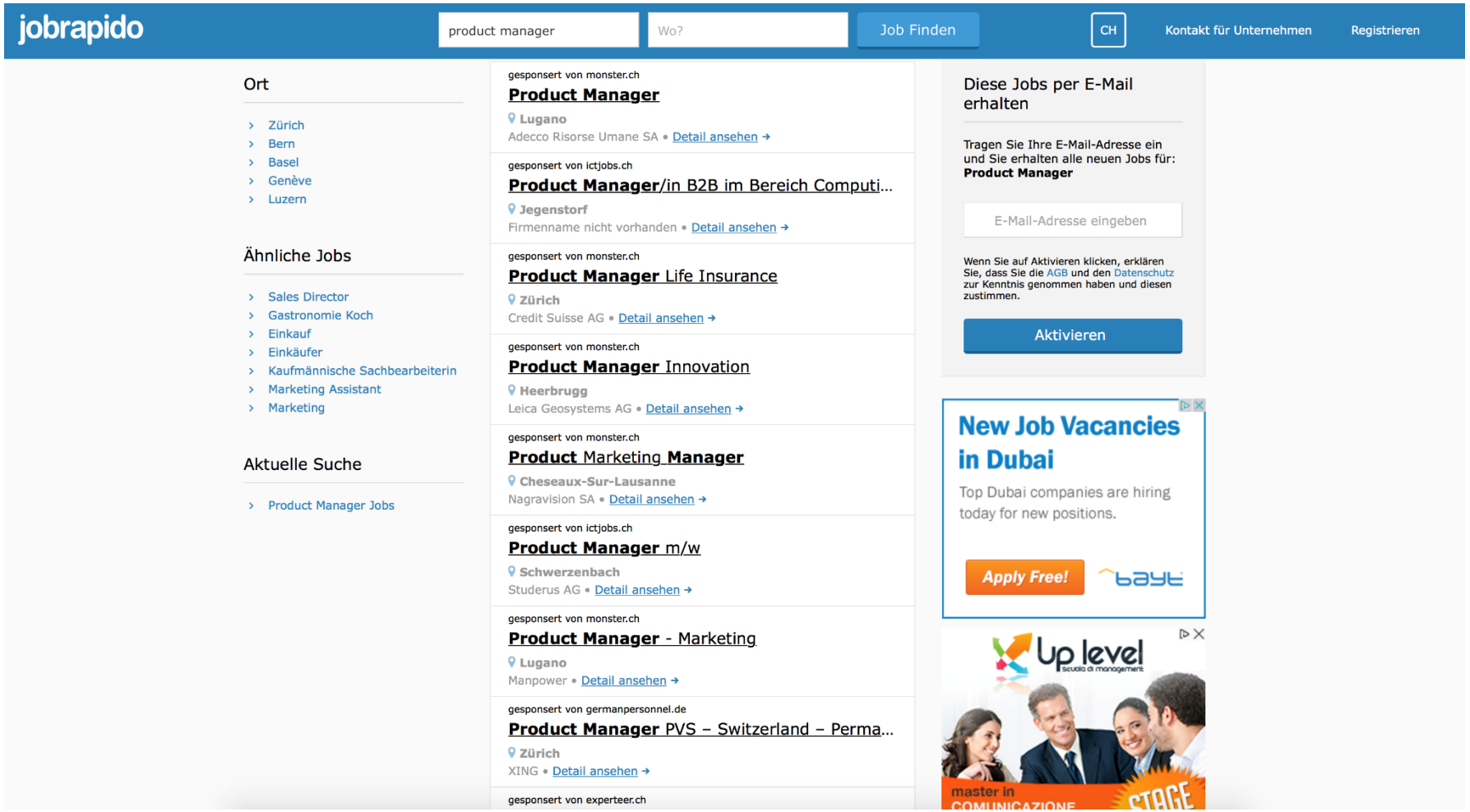
DOCUMENTS
>200 M

CURRENT VERSION
1.7.3

SEACOM
the leading open source architects

jobrapido

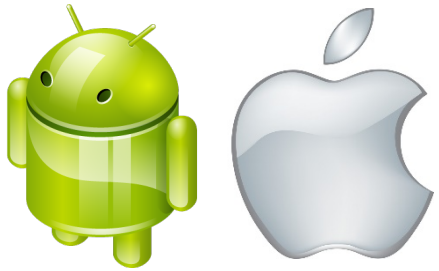
OUR SEARCH PAGE

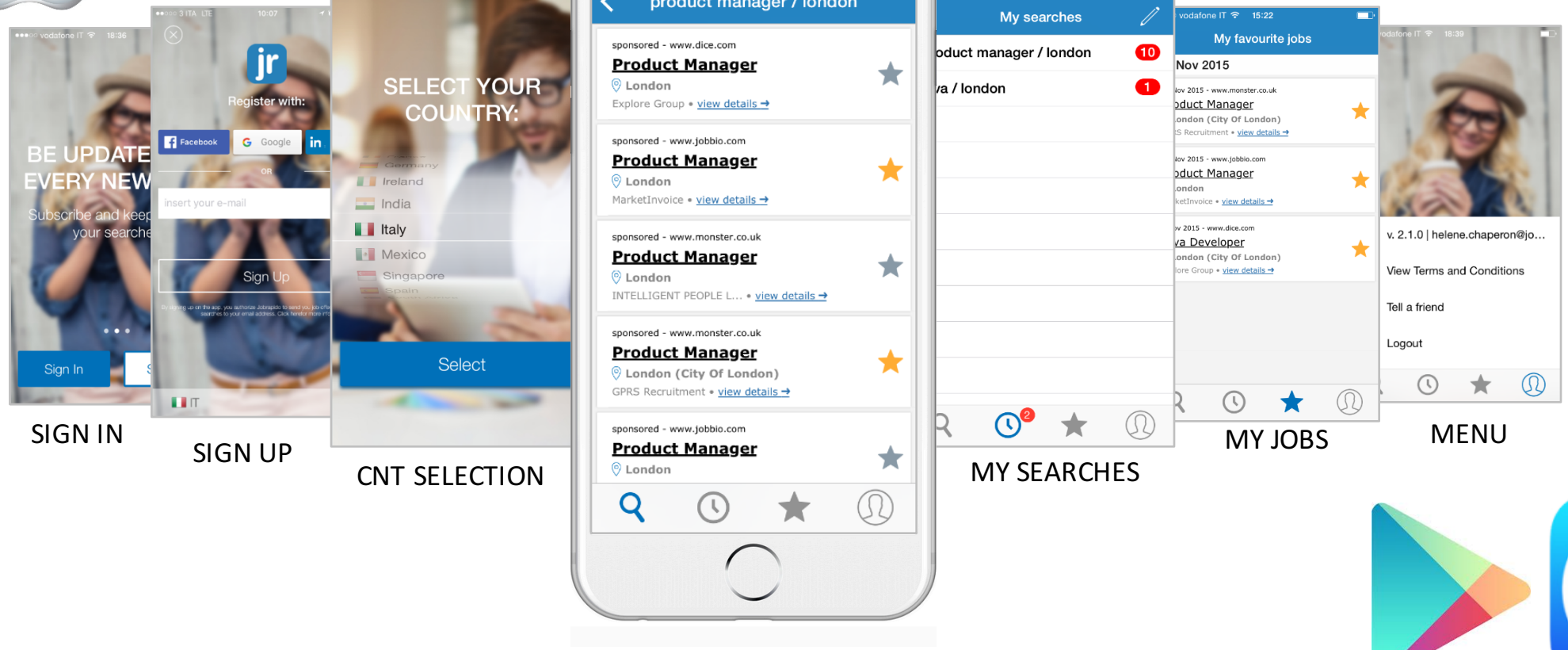


SEACOM
the leading open source architects

jobrapido

MOBILE APP





SIGN IN

SIGN UP

CNT SELECTION

MY SEARCHES

MY JOBS

MENU

SEACOM
the leading open source architects

jobrapido

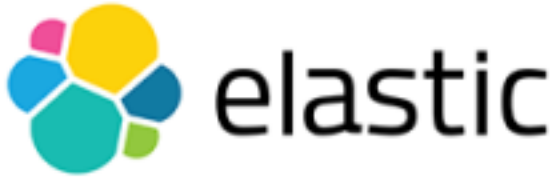
WHERE WE ARE

elastic

SEACOM
the leading open source architects

jobrapido

THE NEED FOR A NEW SEARCH ENGINE



- Result sorting limited to CPC and publish date
- Debug and troubleshooting nearly impossible
- Exact match was the only option
- Slow reindex time (up to 10 days)
- Custom and inaccurate language analysis
- No high availability
- Hard to scale

MULTI-LANGUAGE SUPPORT

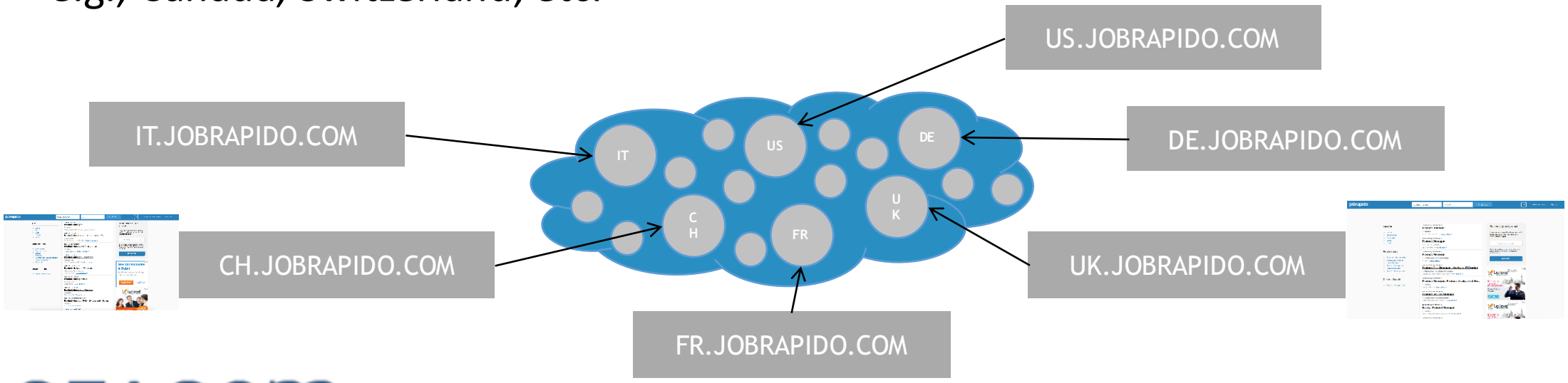
elastic

SEACOM
the leading open source architects

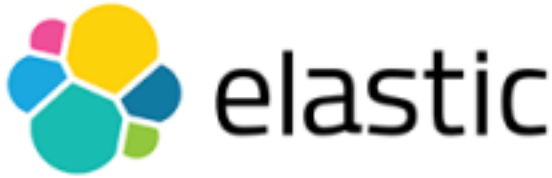
jobrapido

JOB INDICES

- One index per country
  - Two aliases with filter (organic vs. sponsored jobs)
  - Each index implements country and language-specific analysers
- A country may support more than one language
  - e.g., Canada, Switzerland, etc.

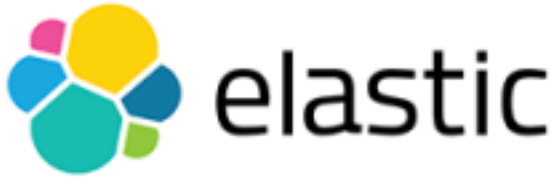


THE ANATOMY OF AN ANALYZER



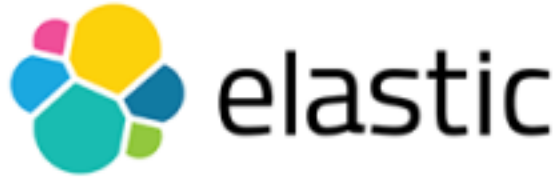
- Strip HTML
- Tokenization
- Lowercase
- Stopwords
  1. _german_, _french, _english_, … (built-in)
  2. Language-specific (file)
  3. Country-specific (file)
- Stemming
  1. light_german, light_french, english, … (built-in)
  2. Language-specific exceptions (file)
  3. Country-specific exceptions (file)
- Language-specific filters (e.g., elision, possessive)
- Synonyms
- Shingles

MULTI-LANGUAGE PROPERTIES

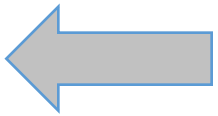


BODY	CHAR FILTER	TOKENIZER	FILTER
GERMAN	HTML STRIP	STANDARD	LOWERCASE, GERMAN STOPWORDS, GERMAN STEMMER
STANDARD	HTML STRIP	STANDARD	STANDARD, LOWERCASE, GERMAN STOPWORDS
ENGLISH	HTML STRIP	STANDARD	POSSESSIVE, LOWERCASE, ENGLISH STOPWORDS, ENGLISH STEMMER
FRENCH	HTML STRIP	STANDARD	ELISION, LOWERCASE, FRENCH STOPWORDS, FRENCH STEMMER
ITALIAN	HTML STRIP	STANDARD	ELISION, LOWERCASE, ITALIAN STOPWORDS, ITALIAN STEMMER
SHINGLE	HTML STRIP	STANDARD	LOWERCASE, GERMAN STOPWORDS, GERMAN STEMMER, SHINGLE

MULTI-LANGUAGE PROPERTIES




```
"query": {
   "filtered": {
      "query": {
         "bool": {
            "must": {
               "multi_match": {
                  "query": "product manager",
                  "fields": [
                     "body^3", //german
                     "body.standard^3”,
                     "body.english^2”,
                     "body.french^1”,
                     "body.italian^1”
                  ],
                  "type": "most_fields",
                  "operator": "AND”
               }
            }
         }
      },
      "filter": { … }
   }
}
```

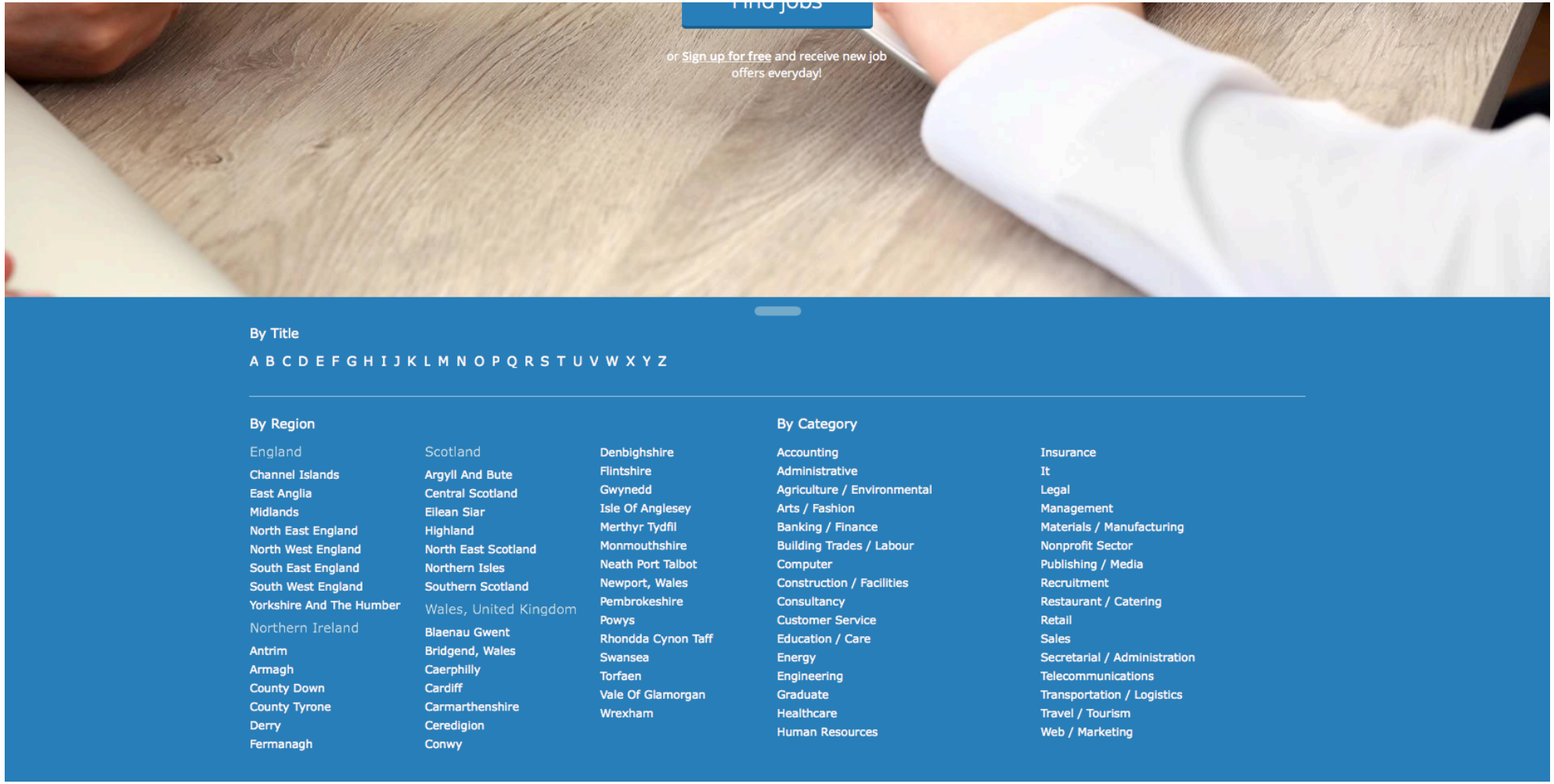


Application-side configurations allow us to define search fields and their individual boost

We constantly run A/B-test to improve matching rate and tune relevance

jobrapido

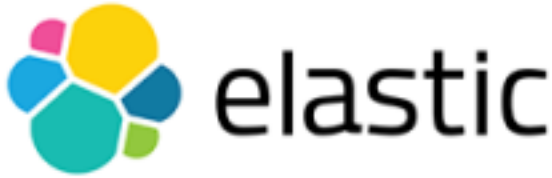
SITEMAPS: PERCOLATORS AND AGGREGATIONS



SEACOM
the leading open source architects

jobrapido

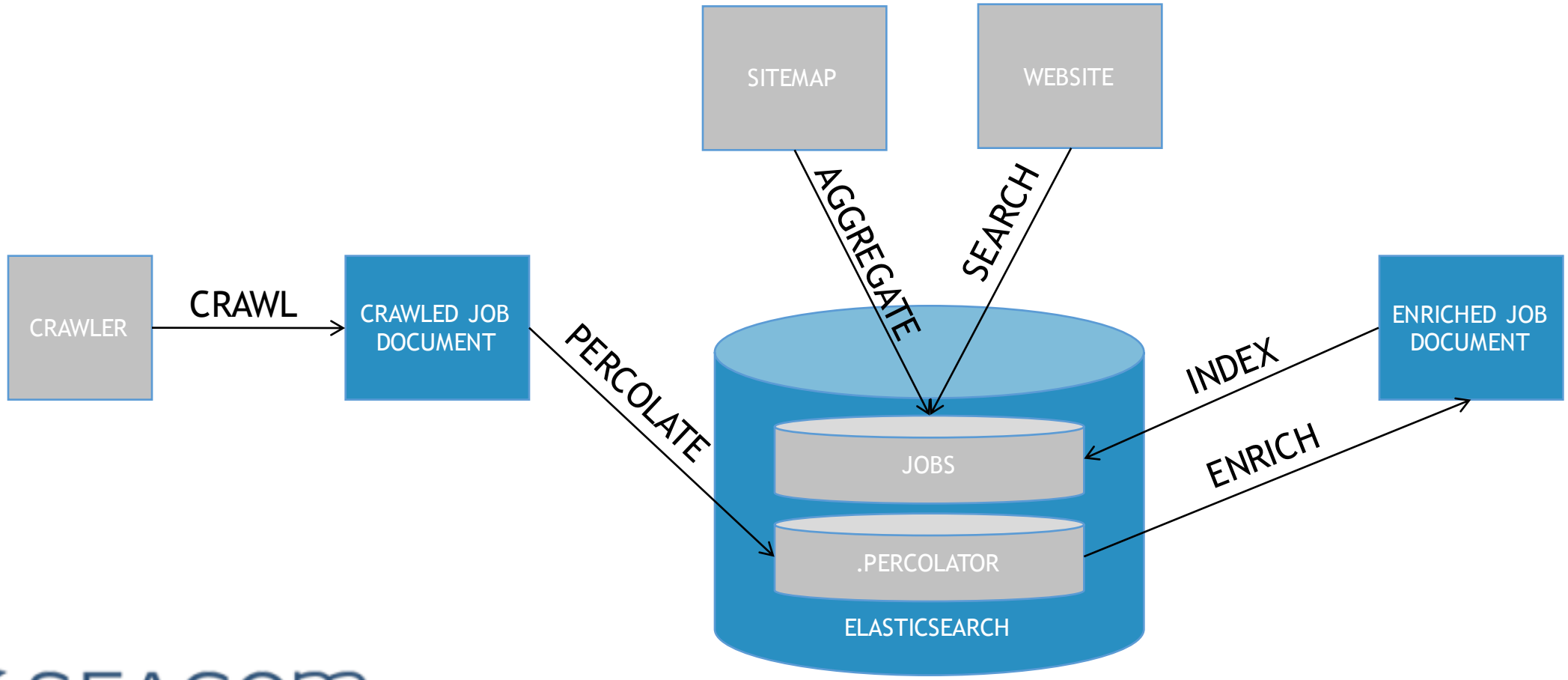
SITEMAP BY JOB TITLES



- Industry is an information you cannot easily find in structured documents
- Only few websites explicitly show job titles and industry

- What if we build a taxonomy of job titles/industry represented by queries?
- That would allow enriching documents at index time by means of percolators

JOURNEY OF A JOB DOCUMENT

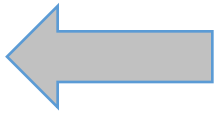


SEACOM
the leading open source architects

jobrapido

PERCOLATOR EXAMPLE

```
"query": {
   "filtered": {
      "query": {
         "bool": {
            "must": {
               "multi_match": {
                  "query": "Account Director",
                  "fields": [
                     "headline^2",
                     "headline.standard^1",
                     "body^1",
                     "body.standard^1",
                     "company_name^1”
                  ],
                  "type": "most_fields",
                  "operator": "AND”
               }
            }
         }
      },
      "filter": {
         "bool": {
            …
         }
      }
   },
   "jobtitle": "Account Director",
   "sector": "Sales"
}
```



Percolator is a standard query (multi match in search fields)

Jobtitle and sector are attached to the query and indexed together with the document (nested)

SEACOM
the leading open source architects

jobrapido

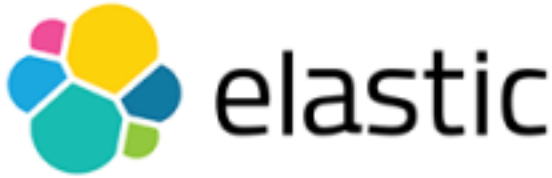
PROS AND CONS

- Live document enrichment (+)
- Job classification based on keywords (+)
- Aggregate by industry and sub-aggregate by location (+)

- Slower reindex time (-)
  - Reindex all 10x slower
- Aggregations are heavy (-)
  - Caching required
  - Inaccurate since the population is dynamic

- Try to be consistent with your queries
  - e.g., percolators do not support min_score, whereas queries do

WHAT'S NEXT

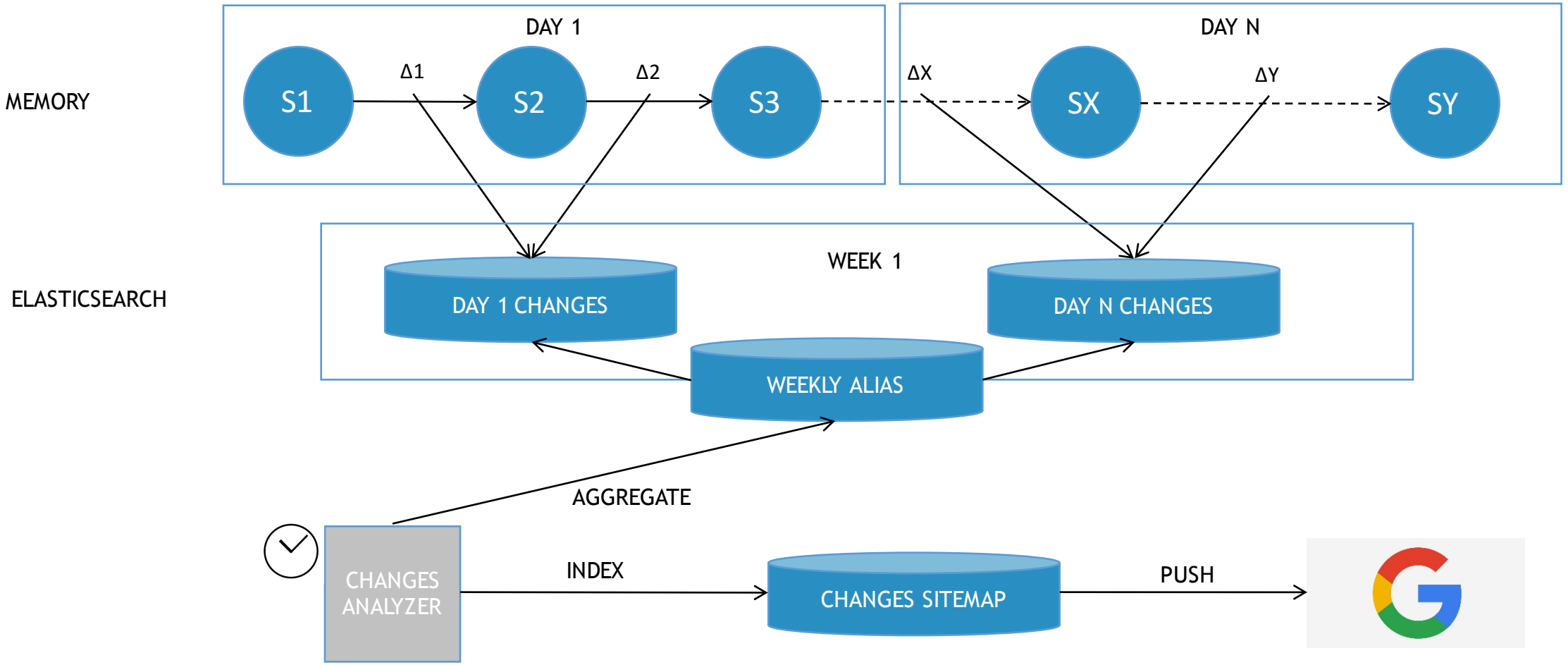


- Sitemaps change frequently
  - Job import and lifecycle cause link churn
- Sitemaps are heavy
  - Tons of jobtitle and locations

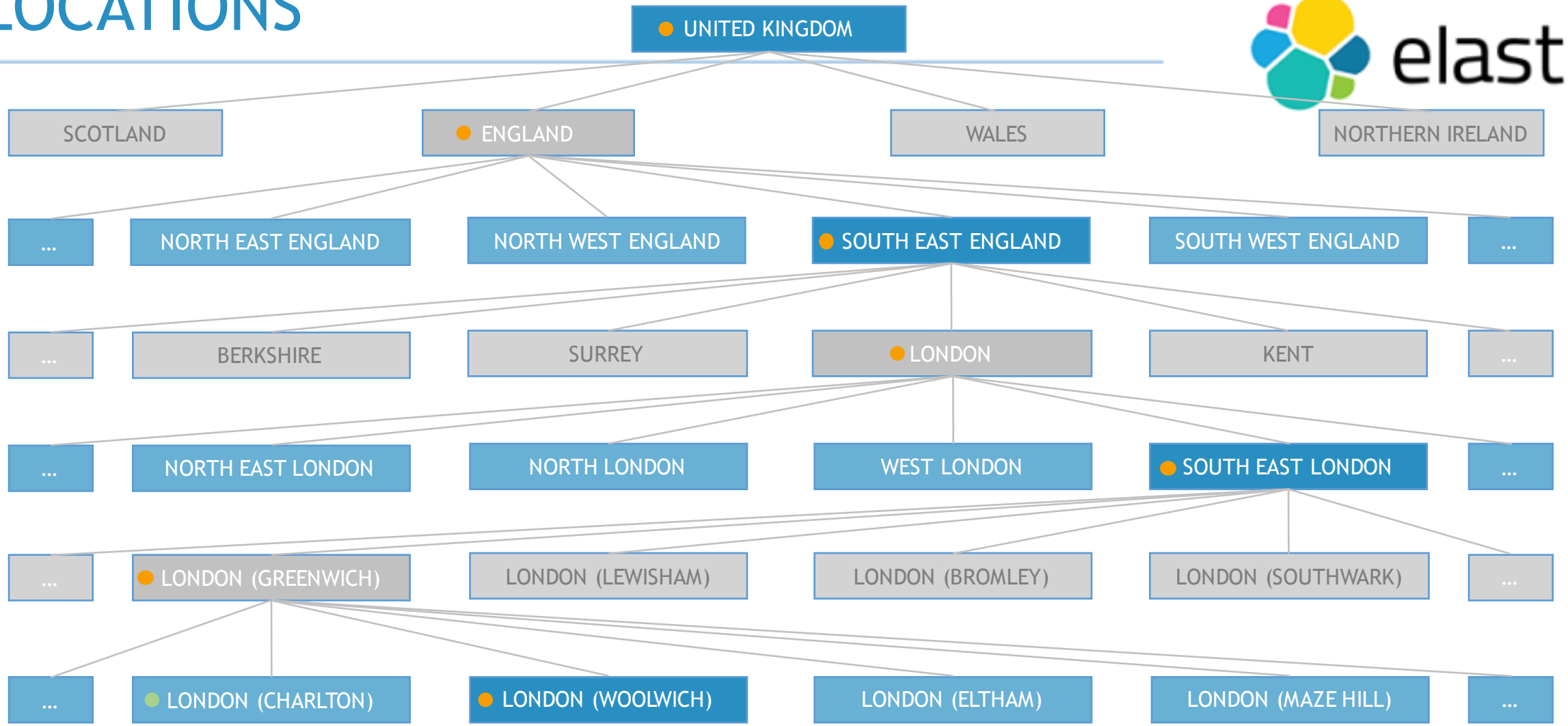
- Google periodically crawls sitemaps
- Google allows pushing sitemap changes
  - We do not want to push unstable changes

jobrapido

SITEMAP CHANGES

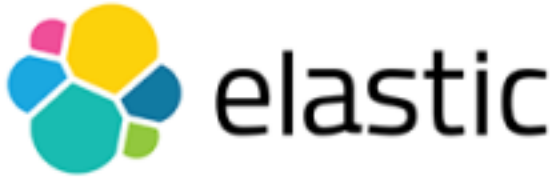


LOCATIONS



/UNITED KINGDOM/ENGLAND/SOUTH EAST ENGLAND/LONDON/SOUTH EAST LONDON/LONDON (GREENWICH)/LONDON (WOOLWICH)
/UNITED KINGDOM/ENGLAND/SOUTH EAST ENGLAND/LONDON/SOUTH EAST LONDON/LONDON (GREENWICH)/LONDON (CHARLTON)

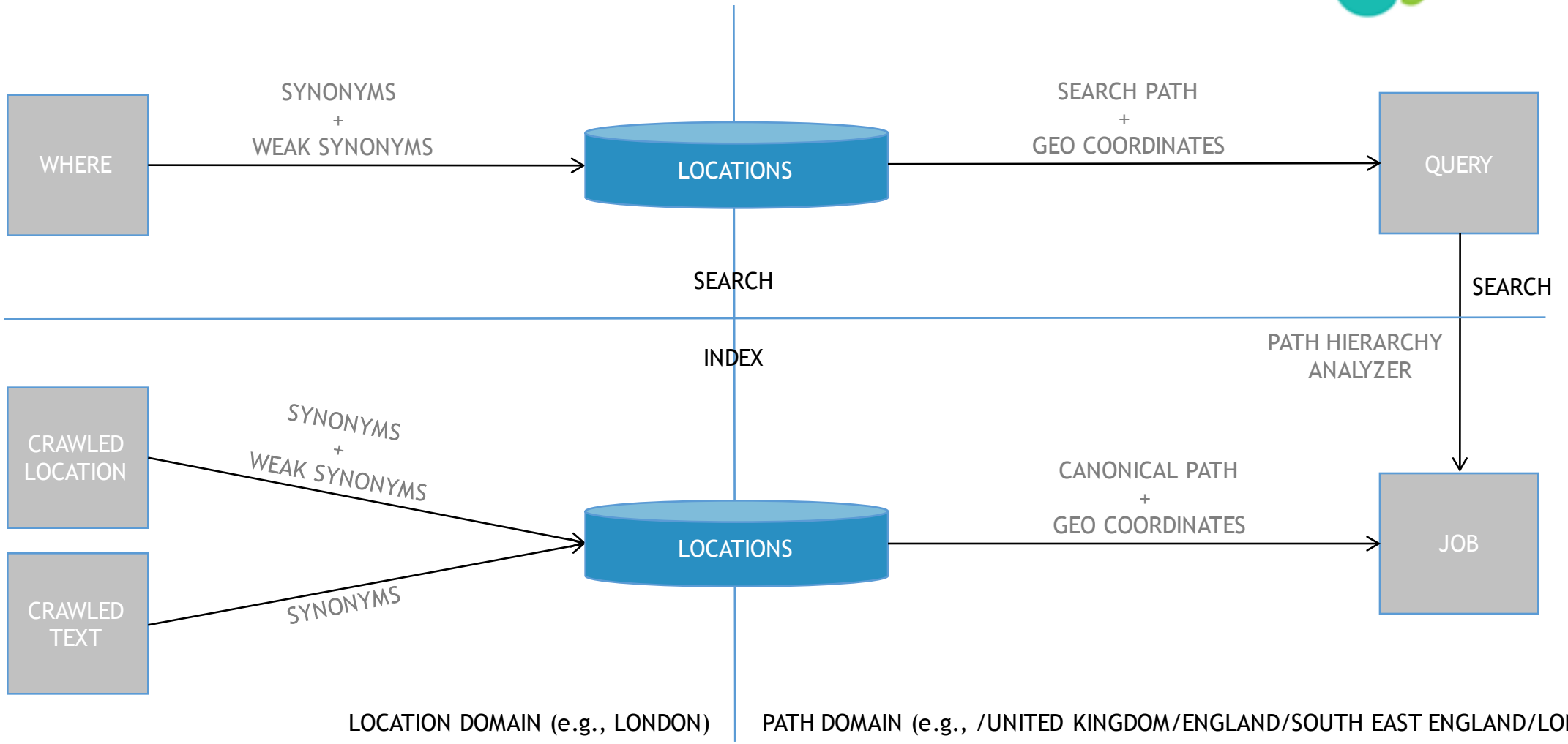
LOCATION MAPPING



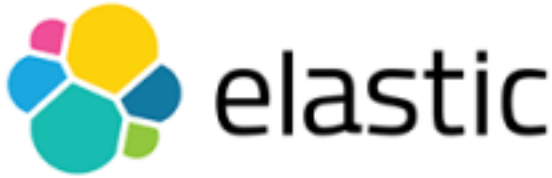
LOCATION	LOCATION
CANONICAL NAME	LONDON
CANONICAL PATH	/UNITED KINGDOM/ENGLAND/SOUTH EAST ENGLAND/LONDON
GEO COORDINATES	POINT (-0.130714000141, 51.498555)
LOCATION DEPTH	3
ORGANIC PATH	/UNITED KINGDOM/ENGLAND/SOUTH EAST ENGLAND/LONDON
SEARCH PATH	{...}
SPECIAL PATH	/LONDON
SYNONYMS	LONDRES, LONDRA, GREATER LONDON, SE1 1PP, EC2A 4JU, ...
WEAK SYNONYMS	[]

jobrapido

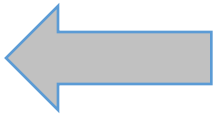
LOCATION SEARCH AND INDEXING



LOCATION SEARCH




```
"or": {
   "filters": [
   {
      "terms": {
         "location": [
            "/hartsville, sc",
            "/united states/southern united states/south atlantic/south carolina/darlington county, sc/hartsville, sc”
         ],
         "_cache": false
      }
   },
   {
      "geo_shape": {
         "geo_coordinates": {
            "indexed_shape": {
               "id": "443498",
               "type": "location",
               "index": "us_geo_shapes",
               "path": "geo_coordinates"
            },
            "relation": "within”
         },
         "_cache": false
      }
   }
   ],
   "_cache": true,
   "_cache_key": "443498"
}
```



We search locations by path and coordinates

Caching is performed only on the or filter (sub-filters always depend on it and we may avoid caching)

Cache keys allow saving memory

CONCLUSIONS

- Jobrapido covers 58 countries and 18 languages
- Percolations and aggregations allow for document enrichment and dynamic sitemap creation
- Pipeline aggregations ease push of significant sitemap changes to Google
- Path hierarchies to cleverly represent location structure
- Index and search like a pro ☺
  - Documentation: https://www.elastic.co/guide/index.html
  - Training: thanks Luca and Karel ☺
  - Book: Elasticsearch – The Definitive Guide
  - Support: Jobrapido is a proud platinum customer (production and development advice) – thanks Antonio ☺

THANK YOU

jobrapido

http://corporate.jobrapido.com

SEACOM
the leading open source architects

Q&A