

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

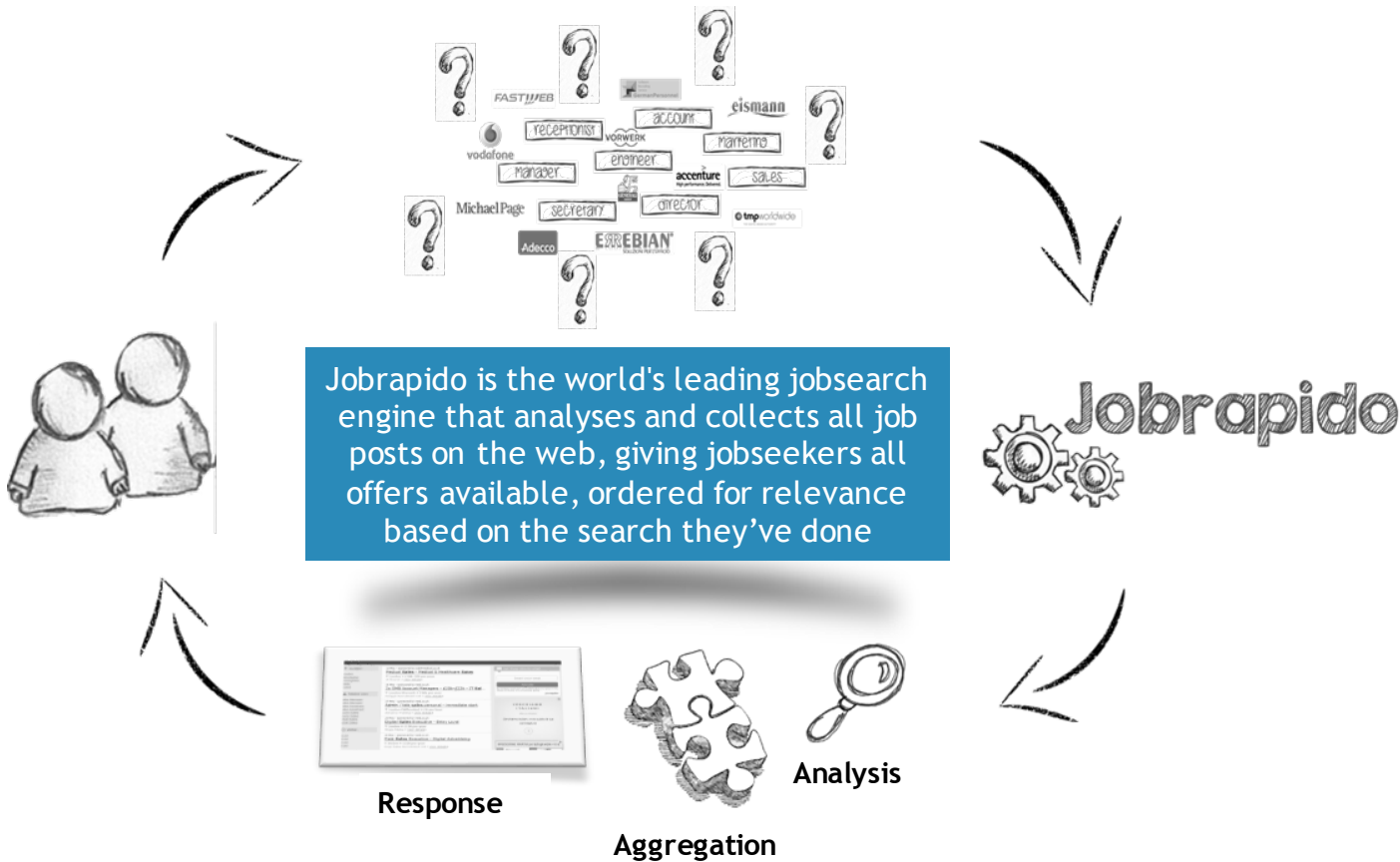
<https://it.linkedin.com/in/salvatorevadacca>

COMPANY WEBSITE

www.jobrapido.com



WHO WE ARE



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

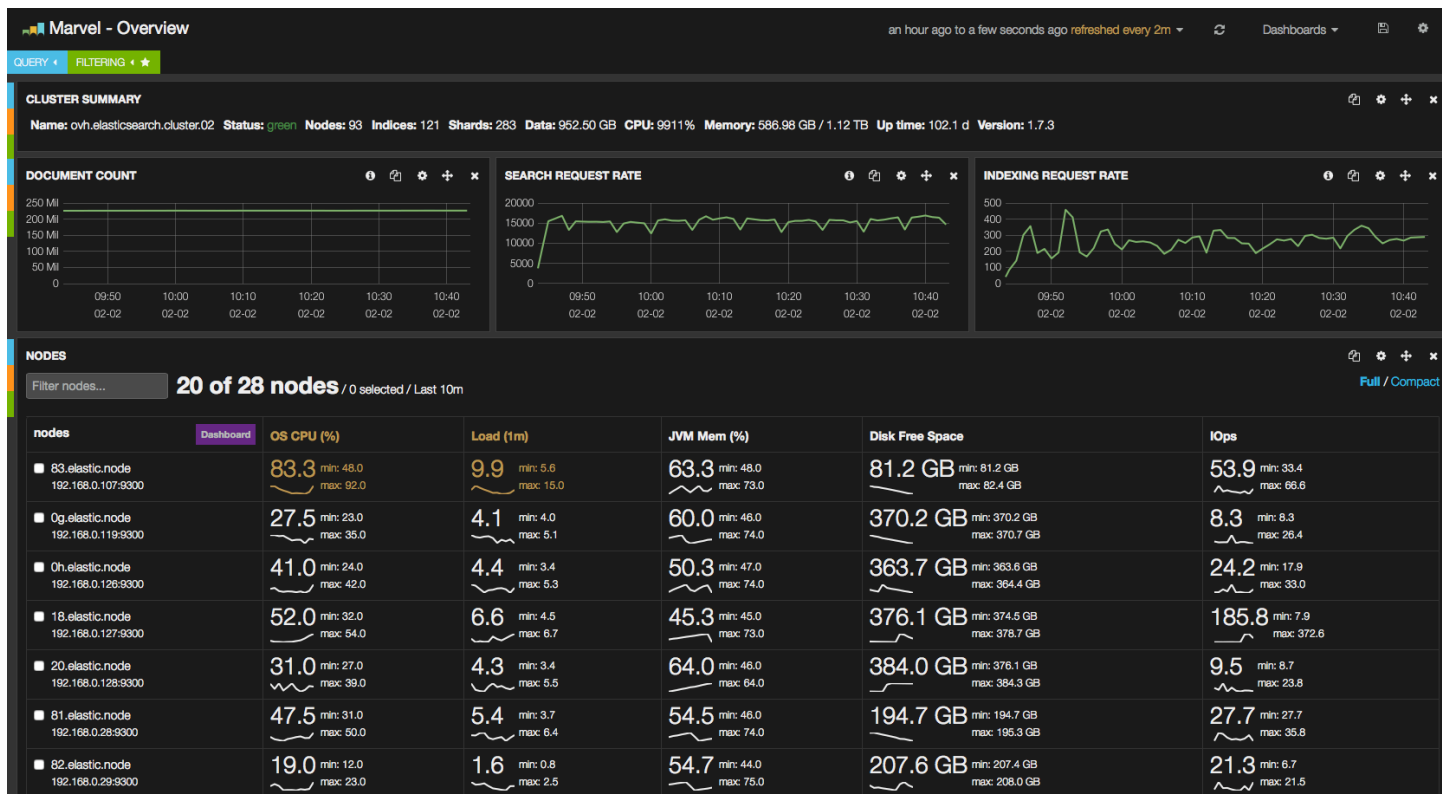
Head office Milan + office in Amsterdam

PEOPLE

100+



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

OUR SEARCH PAGE



jobrapido

product manager Wo? Job Finden CH Kontakt für Unternehmen Registrieren

Ort

- > Zürich
- > Bern
- > Basel
- > Genève
- > Luzern

Ähnliche Jobs

- > Sales Director
- > Gastronomie Koch
- > Einkauf
- > Einkäufer
- > Kaufmännische Sachbearbeiterin
- > Marketing Assistant
- > Marketing

Aktuelle Suche

- > Product Manager Jobs

gesponsert von monster.ch
Product Manager
Lugano
Adecco Risorse Umane SA • [Detail ansehen](#) →

gesponsert von ictjobs.ch
Product Manager/in B2B im Bereich Comput...
Jegenstorf
Firmenname nicht vorhanden • [Detail ansehen](#) →

gesponsert von monster.ch
Product Manager Life Insurance
Zürich
Credit Suisse AG • [Detail ansehen](#) →

gesponsert von monster.ch
Product Manager Innovation
Heerbrugg
Leica Geosystems AG • [Detail ansehen](#) →

gesponsert von monster.ch
Product Marketing Manager
Cheseaux-Sur-Lausanne
Nagravision SA • [Detail ansehen](#) →

gesponsert von ictjobs.ch
Product Manager m/w
Schwerzenbach
Studerus AG • [Detail ansehen](#) →

gesponsert von monster.ch
Product Manager - Marketing
Lugano
Manpower • [Detail ansehen](#) →

gesponsert von germanpersonnel.de
Product Manager PVS – Switzerland – Perma...
Zürich
XING • [Detail ansehen](#) →

gesponsert von experteer.ch

Diese Jobs per E-Mail erhalten

Tragen Sie Ihre E-Mail-Adresse ein und Sie erhalten alle neuen Jobs für: **Product Manager**

E-Mail-Adresse eingeben

Wenn Sie auf Aktivieren klicken, erklären Sie, dass Sie die AGB und den Datenschutz zur Kenntnis genommen haben und diesen zustimmen.

Aktivieren

New Job Vacancies in Dubai

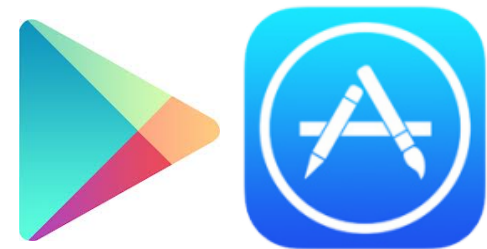
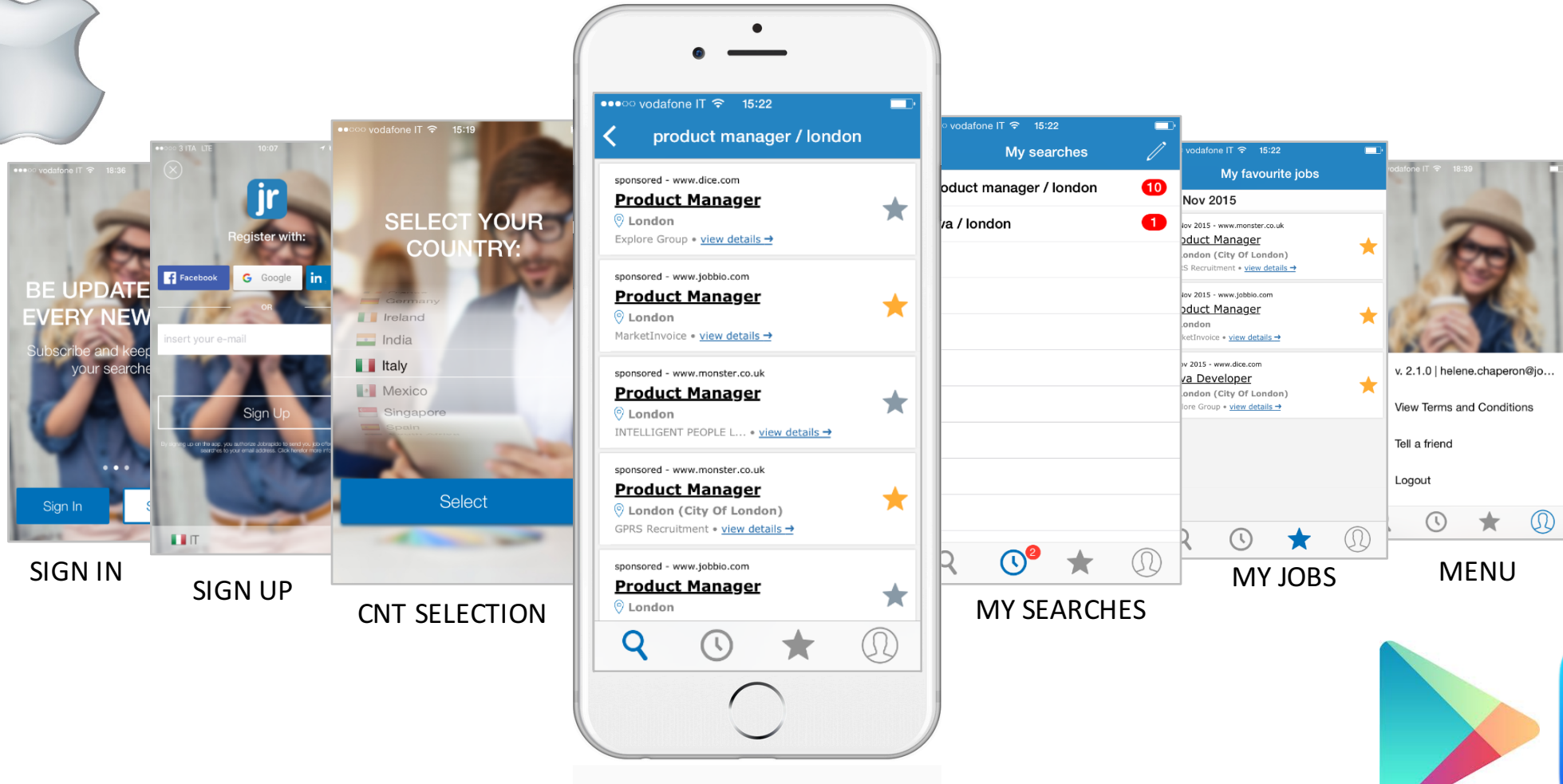
Top Dubai companies are hiring today for new positions.

Apply Free! bayt

Up level
Studio di management

master in COMUNICAZIONE STAGE

MOBILE APP



WHERE WE ARE



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

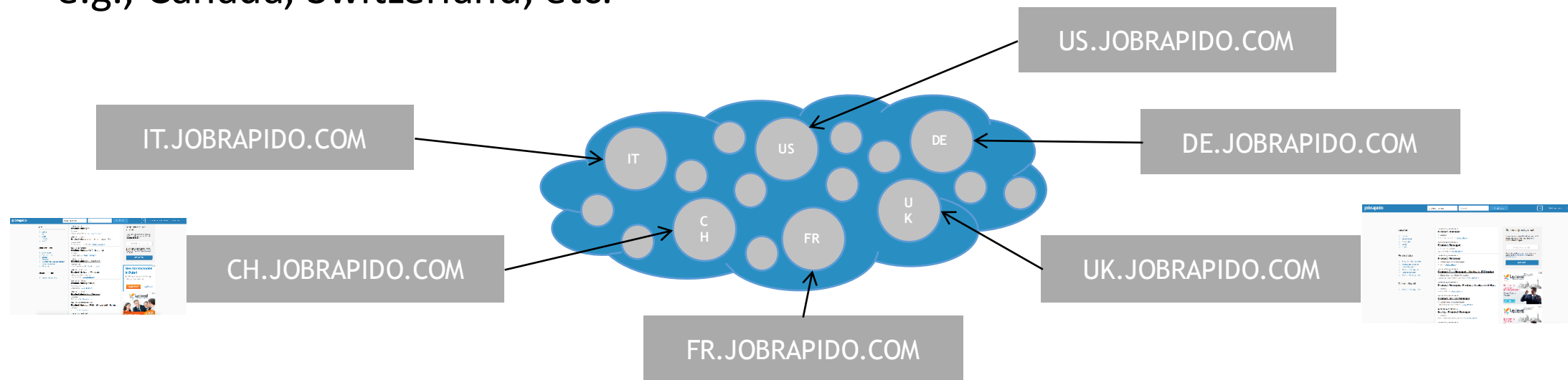


ITALIAN SWEDISH CHINESE
JAPANESE FRENCH HUNGARIAN POLISH CZECH
GERMAN SPANISH RUSSIAN TURKISH
DANISH DUTCH PORTUGUESE
ENGLISH ROMANIAN KOREAN

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

SITEMAPS: PERCOLATORS AND AGGREGATIONS



Find jobs

or Sign up for free and receive new job offers everyday!

By Title

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

By Region

England	Scotland	Denbighshire
Channel Islands	Argyll And Bute	Flintshire
East Anglia	Central Scotland	Gwynedd
Midlands	Eilean Siar	Isle Of Anglesey
North East England	Highland	Merthyr Tydfil
North West England	North East Scotland	Monmouthshire
South East England	Northern Isles	Neath Port Talbot
South West England	Southern Scotland	Newport, Wales
Yorkshire And The Humber	Wales, United Kingdom	Pembrokeshire
Northern Ireland	Blaenau Gwent	Powys
Antrim	Bridgend, Wales	Rhondda Cynon Taff
Armagh	Caerphilly	Swansea
County Down	Cardiff	Torfaen
County Tyrone	Carmarthenshire	Vale Of Glamorgan
Derry	Ceredigion	Wrexham
Fermanagh	Conwy	

By Category

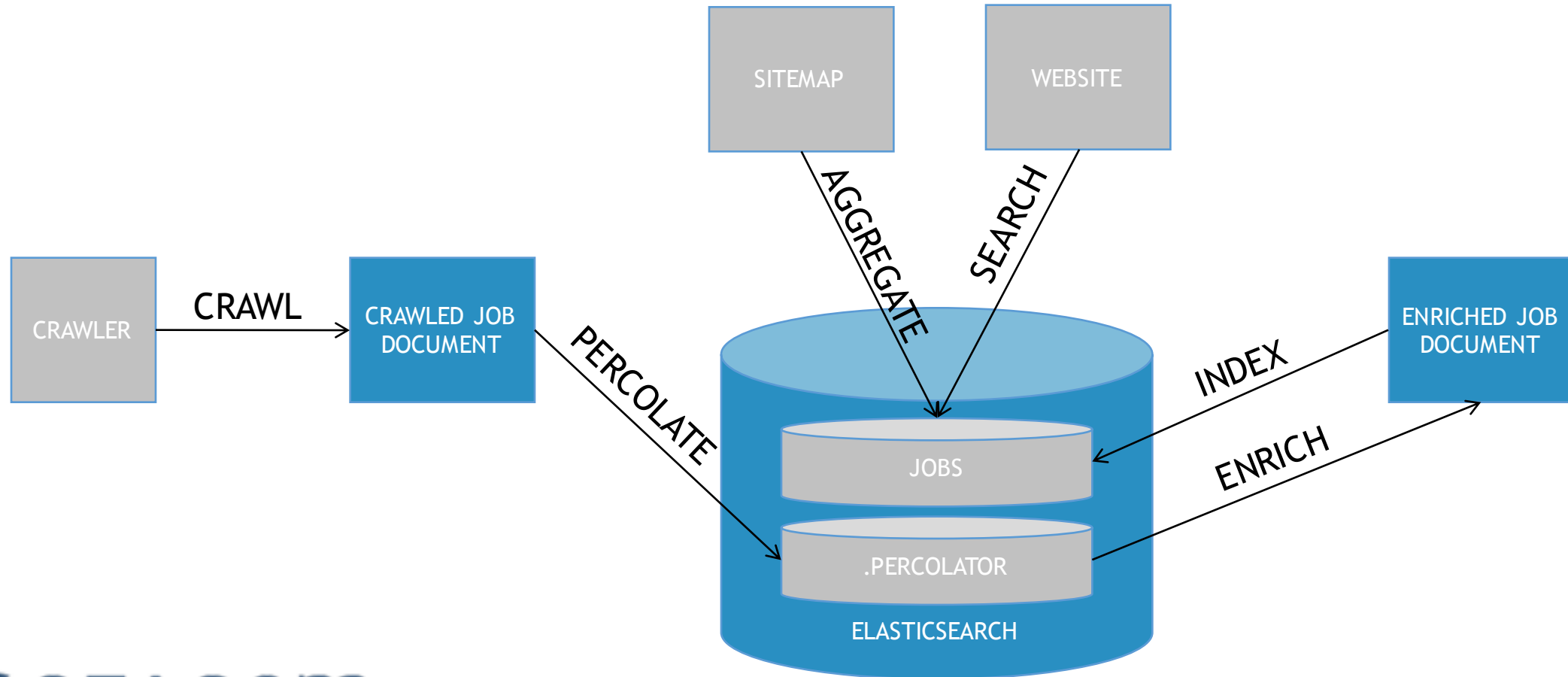
Accounting	Insurance
Administrative	It
Agriculture / Environmental	Legal
Arts / Fashion	Management
Banking / Finance	Materials / Manufacturing
Building Trades / Labour	Nonprofit Sector
Computer	Publishing / Media
Construction / Facilities	Recruitment
Consultancy	Restaurant / Catering
Customer Service	Retail
Education / Care	Sales
Energy	Secretarial / Administration
Engineering	Telecommunications
Graduate	Transportation / Logistics
Healthcare	Travel / Tourism
Human Resources	Web / Marketing

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

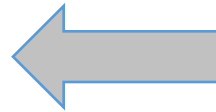
JOURNEY OF A JOB DOCUMENT



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)

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

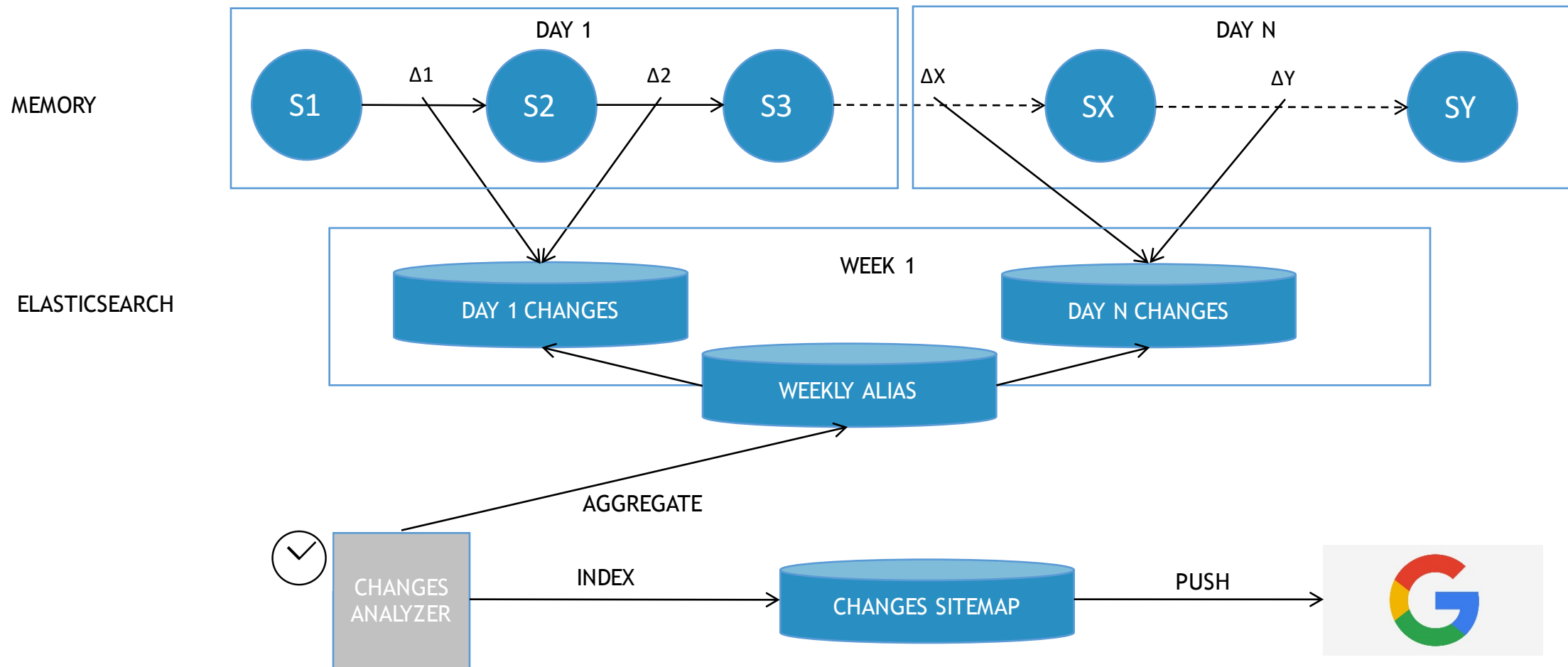
- Try to be consistent with your queries
 - e.g., percolators do not support min_score, whereas queries do

WHAT'S NEXT

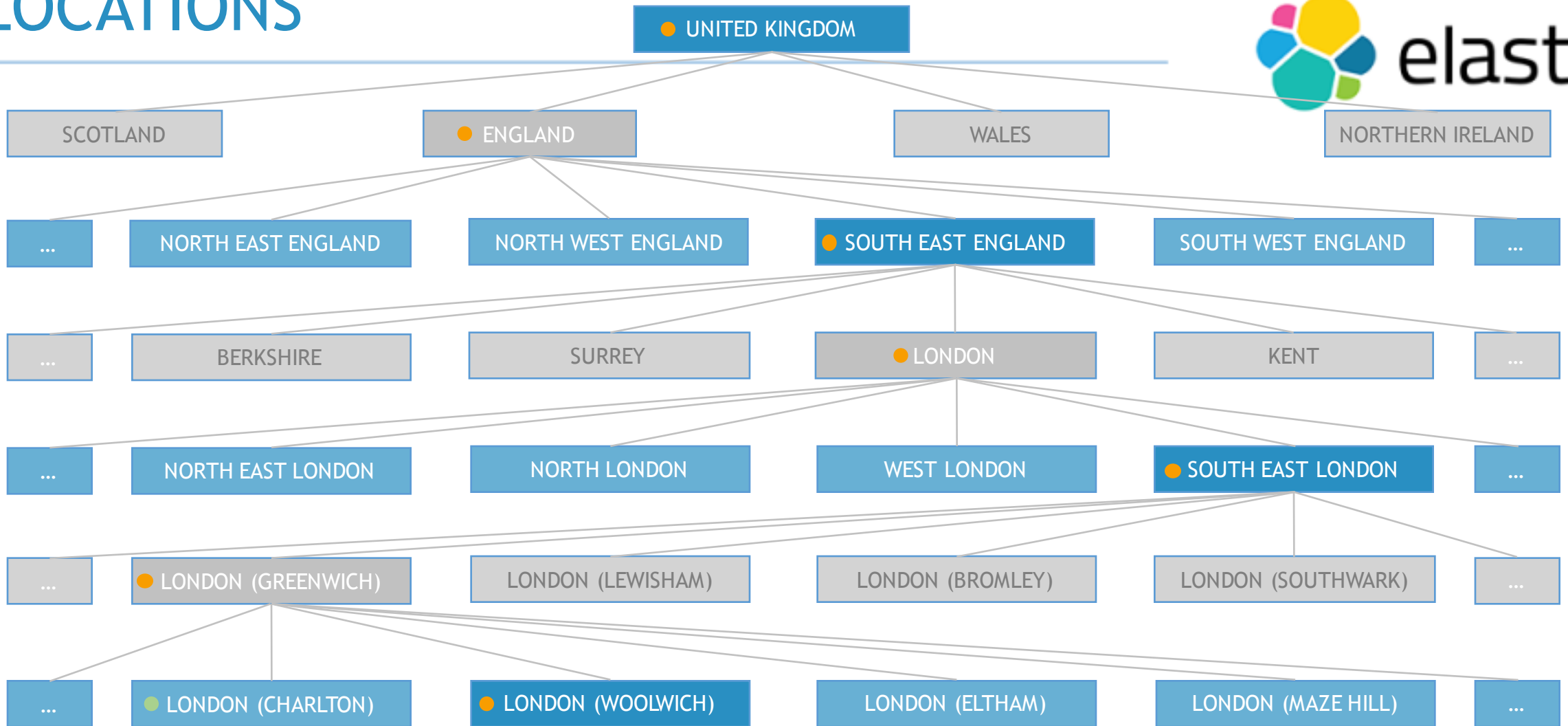
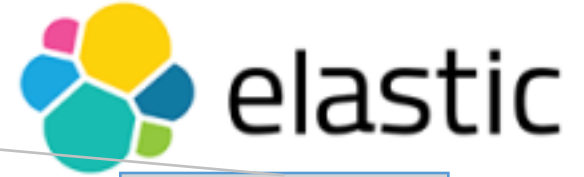


- Sitemaps change frequently
 - Job import and lifecycle cause link churn
- Sitemaps are heavy
 - Tons of jobtitle and locations
- Google periodically crawls sitemaps
- Google allows pushing sitemap changes
 - We do not want to push unstable changes

SITEMAP CHANGES



LOCATIONS



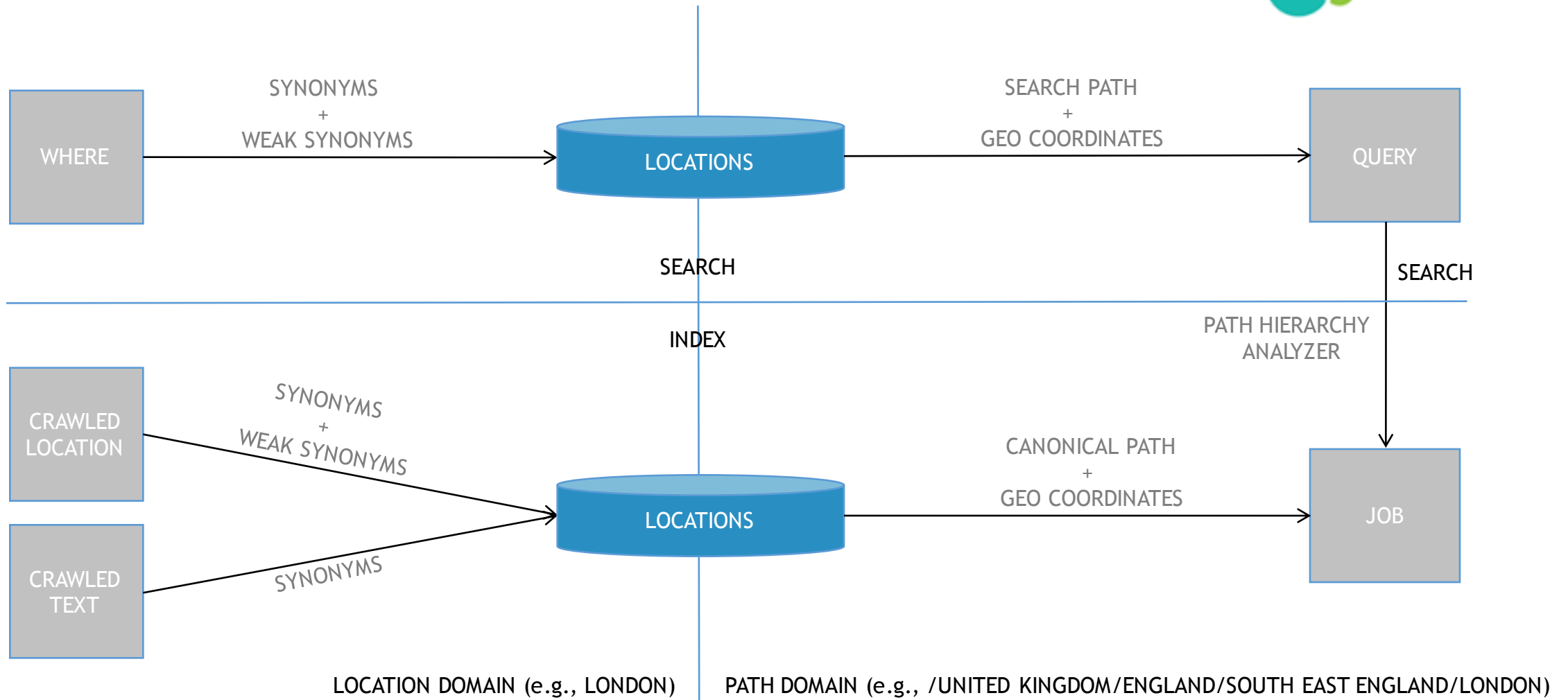
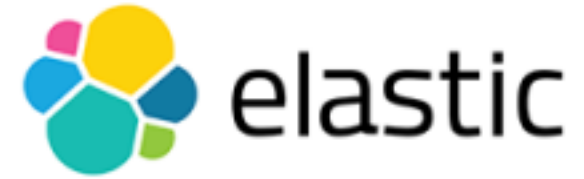
/UNITED KINGDOM/ENGLAND/SOUTH EAST ENGLAND/LONDON/SOUTH EAST LONDON/LONDON (GREENWICH)/LONDON (WOOLWICH)
/UNITED KINGDOM/ENGLAND/SOUTH EAST ENGLAND/LONDON/SOUTH EAST LONDON/LONDON (GREENWICH)/LONDON (CHARLTON)

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	[]

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>

Q&A