Elasticsearch vs Apache Solr

...

Battle of the Open-Source Search Giants !

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Getting acquainted - Who am I?

- Valentin Crettaz (mailto:valentin.crettaz@consulthys.com)
  - https://www.linkedin.com/in/valentincrettaz/
  - https://twitter.com/consulthys
- Developing Java software since 1996
- Fell in love with Elasticsearch in 2010 (v0.9.0)
  - https://www.elastic.co/blog/you-know-for-search
- Running Elasticsearch meetups in Switzerland since early 2016
  - https://www.meetup.com/fr-FR/elasticsearch-switzerland
- Active open-source contributor
  - https://github.com/consulthys
- Active Stack Overflow contributor
  - http://stackoverflow.com/users/4604579/val

Getting acquainted - Who are you?

- How many of you have already...
  - ... heard of Elasticsearch?
  - ... downloaded Elasticsearch?
  - ... installed/run Elasticsearch?
  - ... extended Elasticsearch?

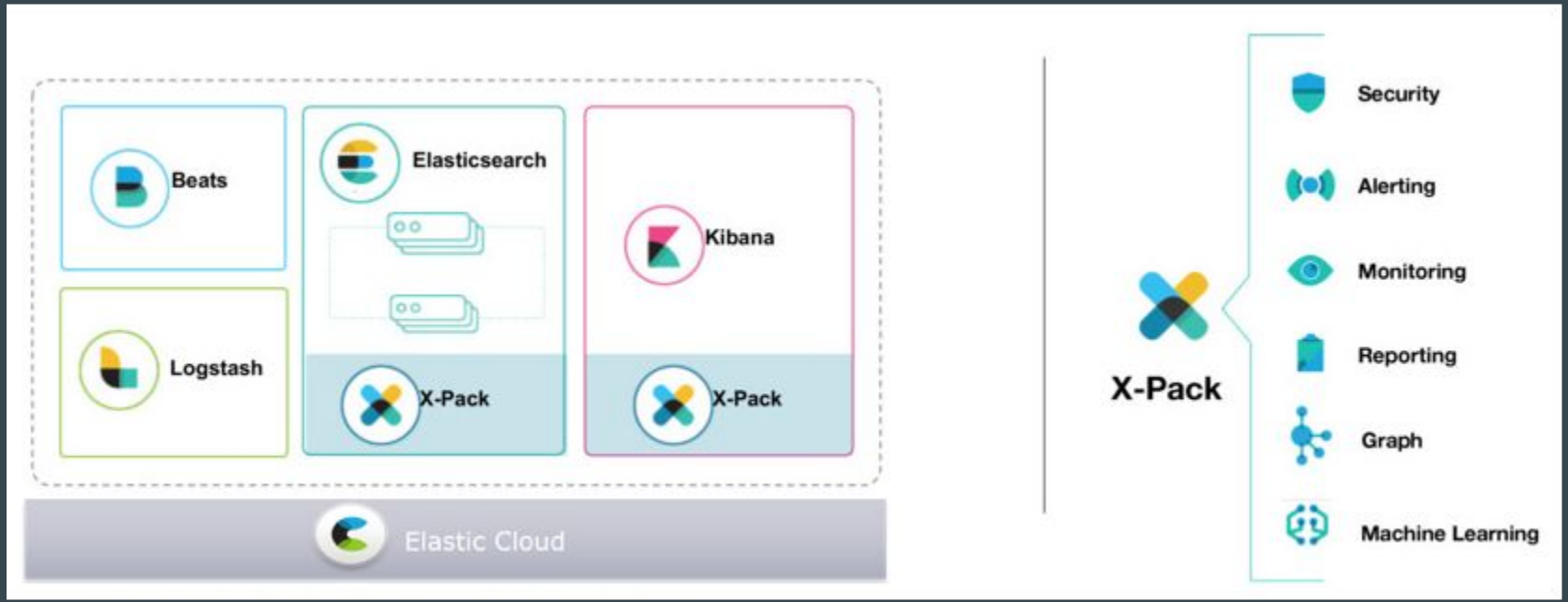
- Your background?

- Define Elasticsearch with your own words

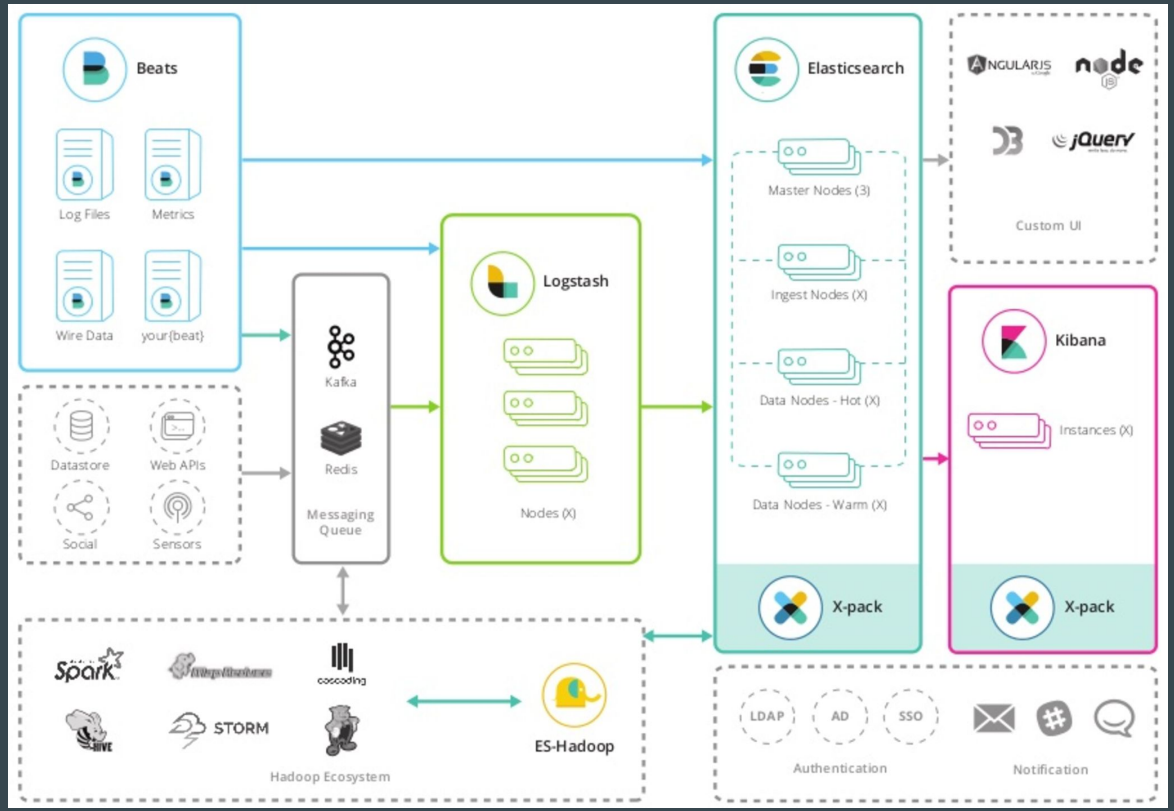
Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

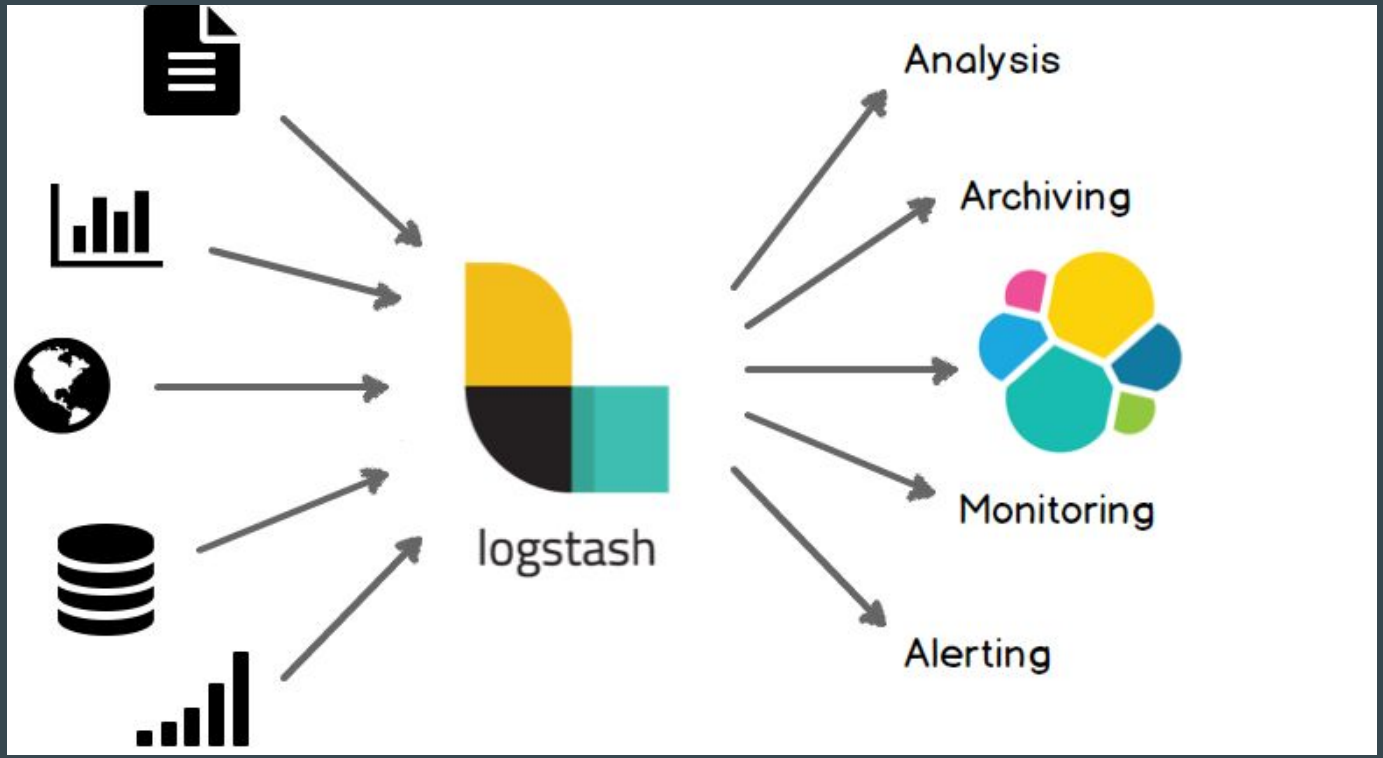
Ecosystem - The Big Picture



Ecosystem - An Even Bigger Picture

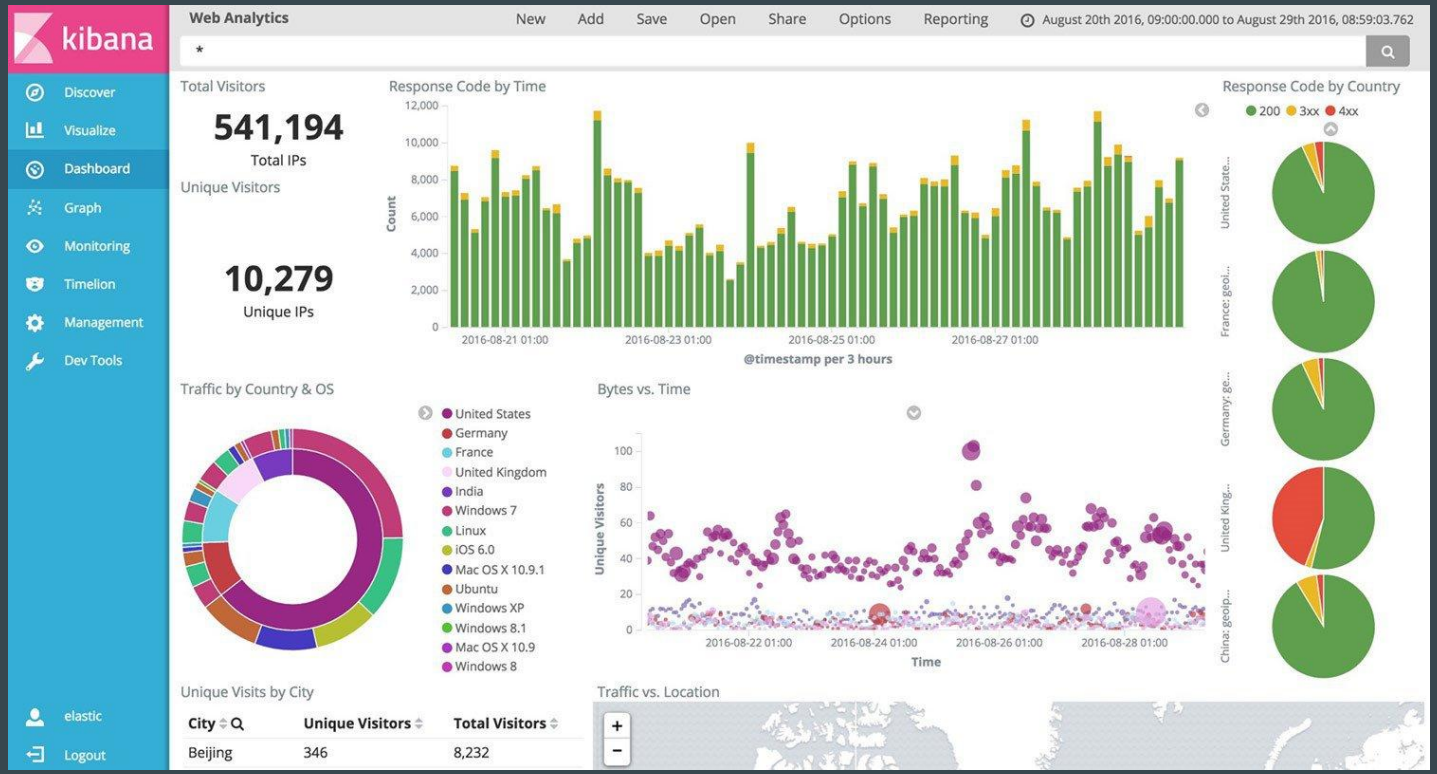


Ecosystem - Logstash



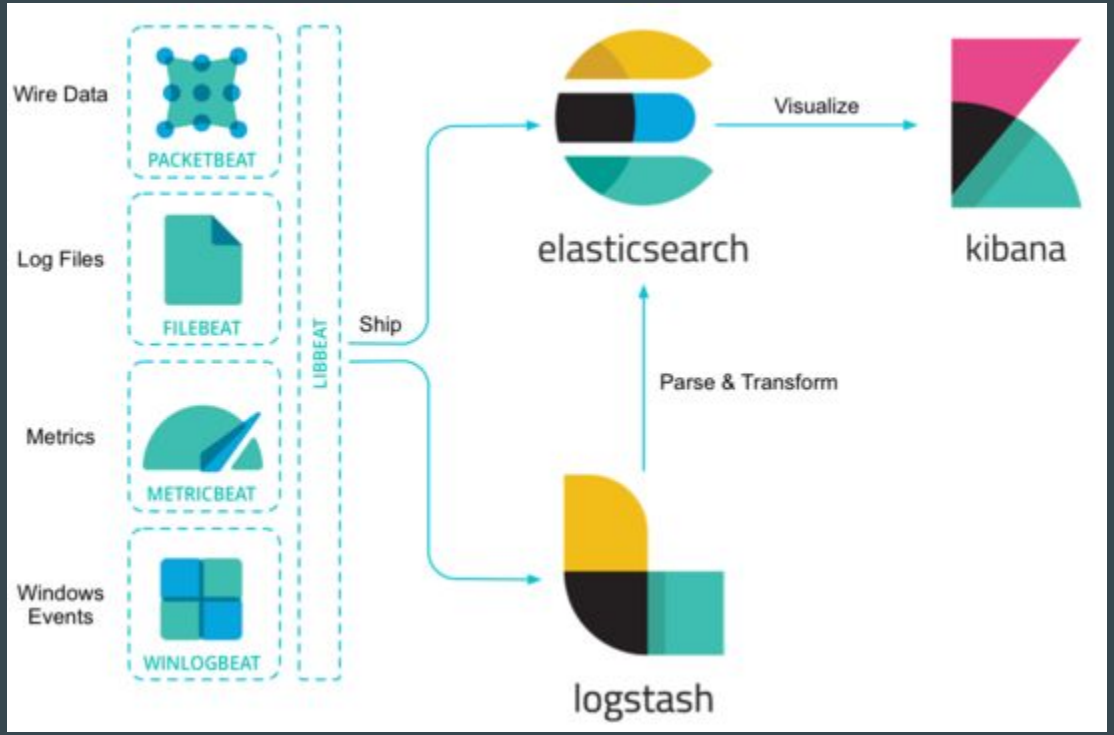
2013: https://www.elastic.co/blog/welcome-jordan-logstash

Ecosystem - Kibana



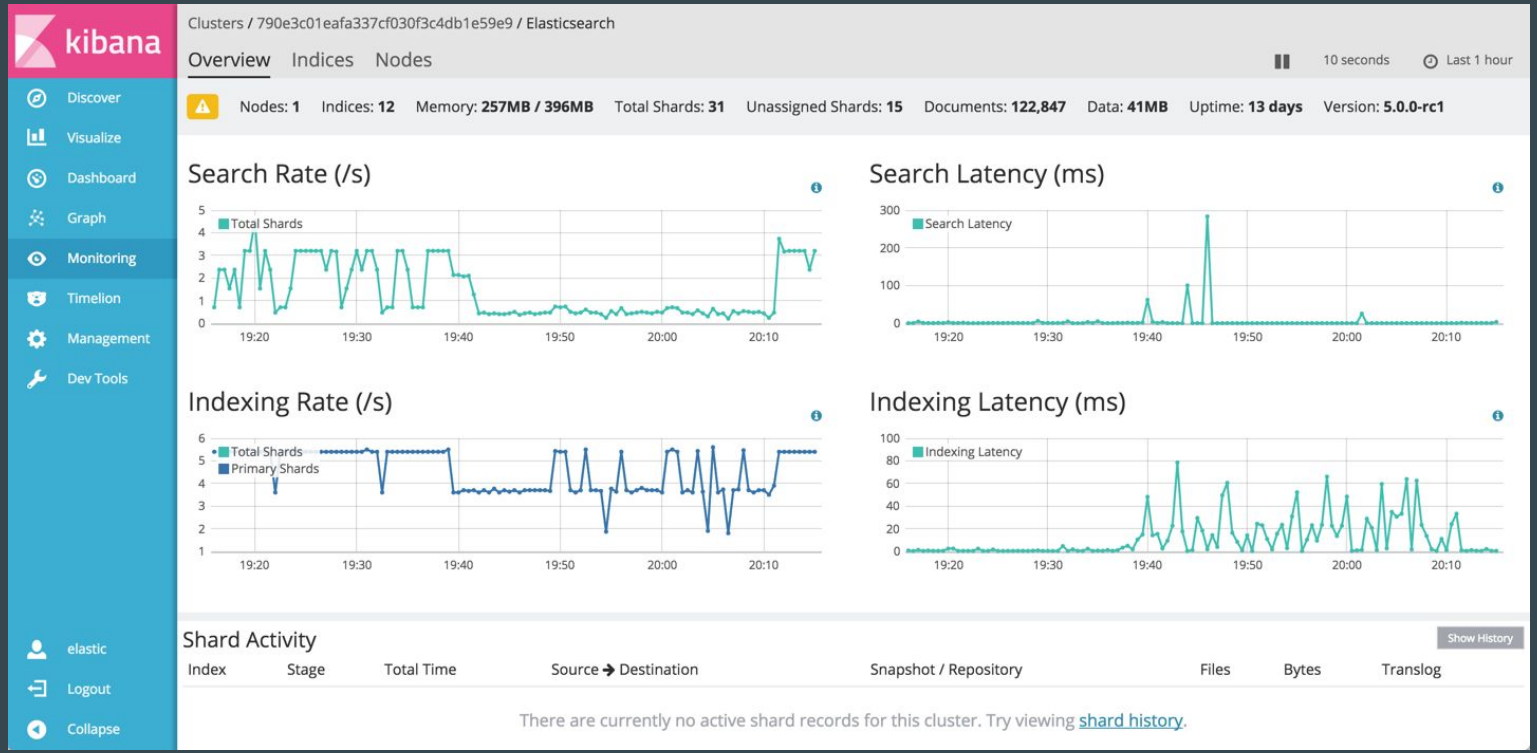
2013: https://www.elastic.co/blog/welcome-drew-rashid

Ecosystem - Beats

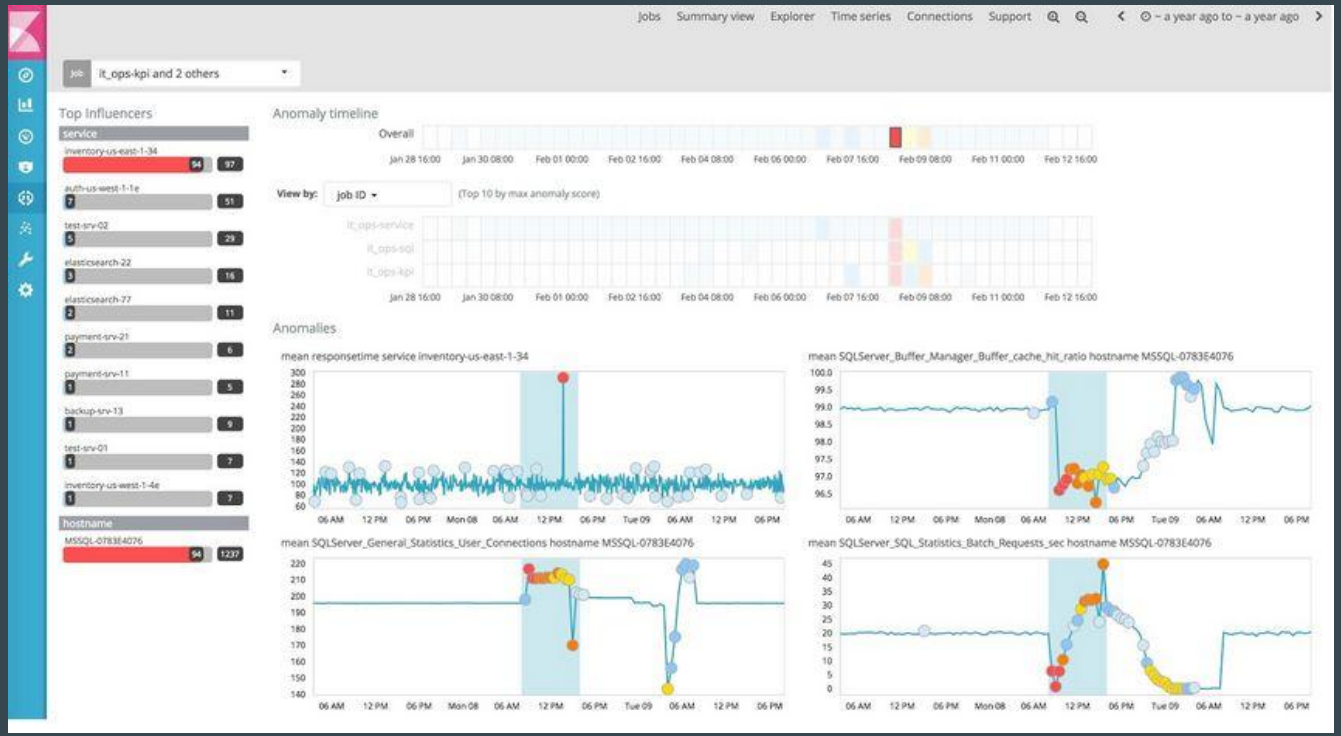


2015: https://www.elastic.co/blog/welcome-packetbeat-tudor-monica

Ecosystem - Monitoring

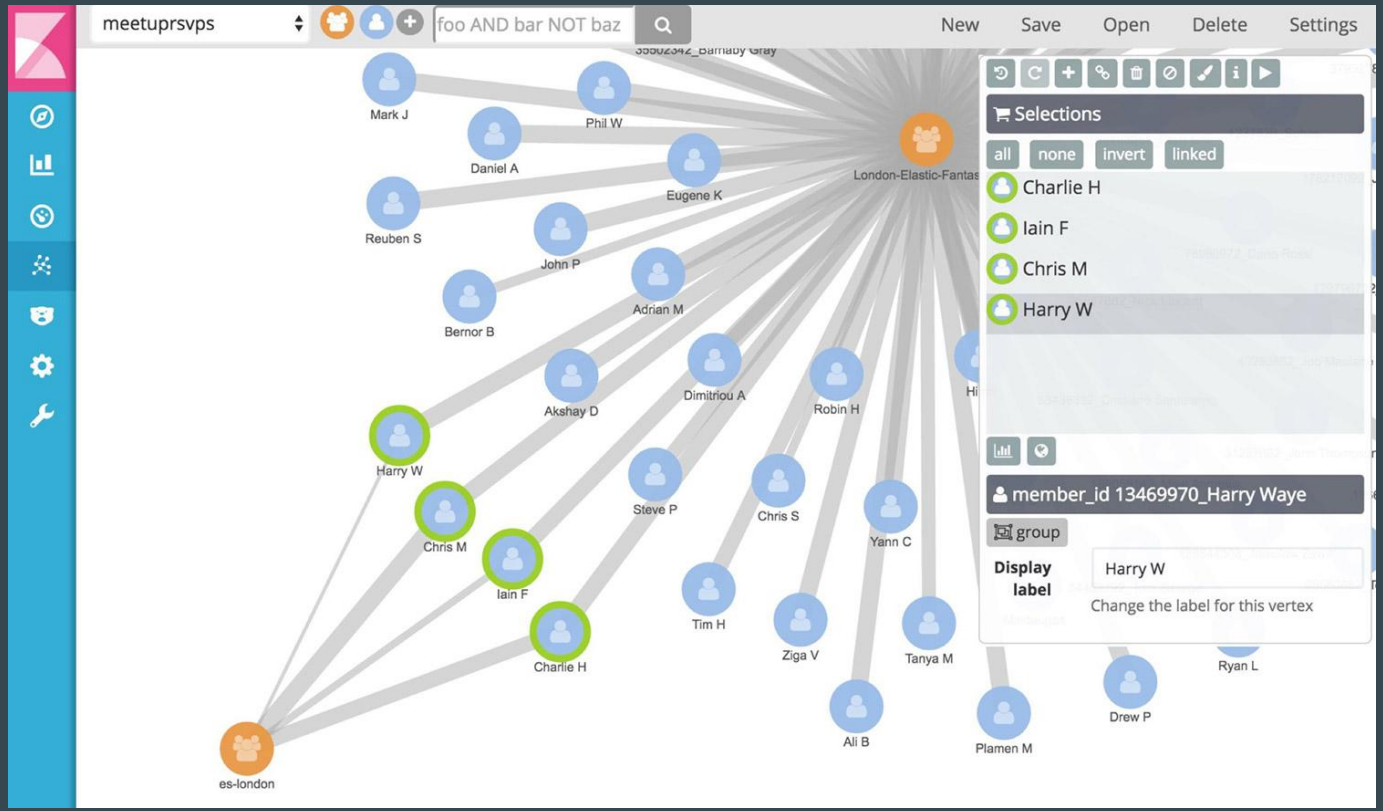


Ecosystem - Machine Learning

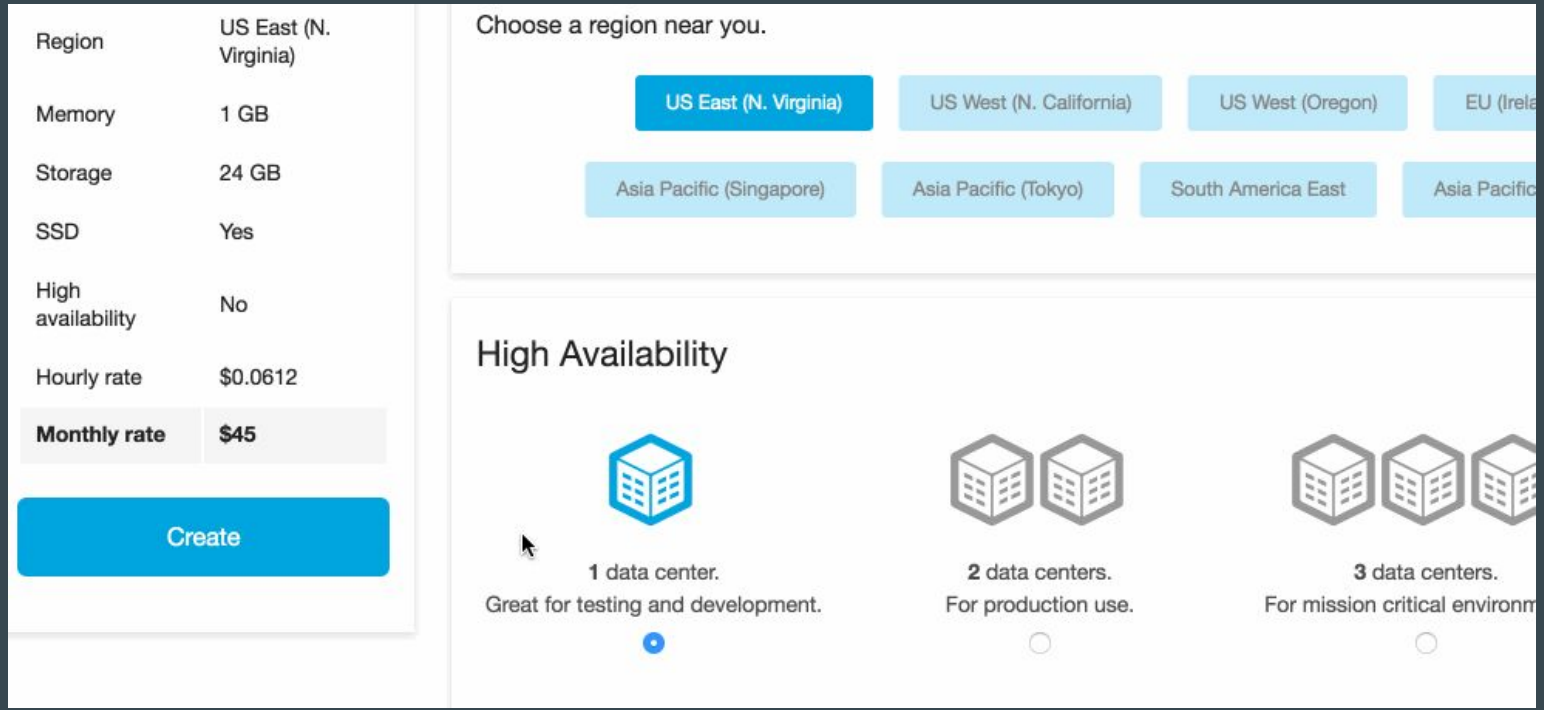


2016: https://www.elastic.co/blog/welcome-prelert-to-the-elastic-team

Ecosystem - Graph

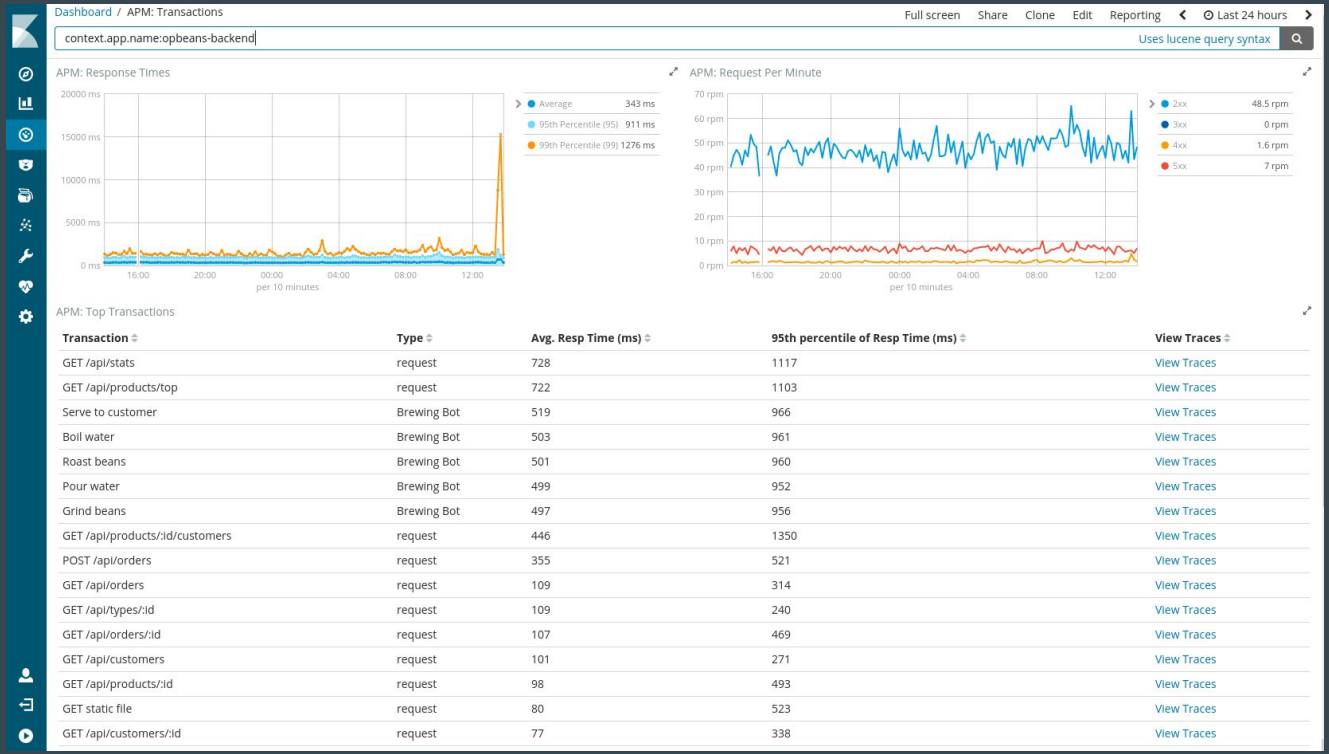


Ecosystem - Elastic Cloud



2015: https://www.elastic.co/blog/welcome-found

Ecosystem - APM



Transaction	Type	Avg. Resp Time (ms)	95th percentile of Resp Time (ms)	View Traces
GET /api/stats	request	728	1117	View Traces
GET /api/products/top	request	722	1103	View Traces
Serve to customer	Brewing Bot	519	966	View Traces
Boil water	Brewing Bot	503	961	View Traces
Roast beans	Brewing Bot	501	960	View Traces
Pour water	Brewing Bot	499	952	View Traces
Grind beans	Brewing Bot	497	956	View Traces
GET /api/products/:id/customers	request	446	1350	View Traces
POST /api/orders	request	355	521	View Traces
GET /api/orders	request	109	314	View Traces
GET /api/types/:id	request	109	240	View Traces
GET /api/orders/:id	request	107	469	View Traces
GET /api/customers	request	101	271	View Traces
GET /api/products/:id	request	98	493	View Traces
GET static file	request	80	523	View Traces
GET /api/customers/:id	request	77	338	View Traces

2017: https://www.elastic.co/blog/welcome-opbeat-to-the-elastic-family

Ecosystem - Site + Entreprise Search

2017: https://www.elastic.co/blog/swiftype-joins-forces-with-elastic

Ecosystem - and more...

- Watcher (for alerting)
- Shield (for security)
- Report (for reporting)
- Canvas (for infographics)
- ...

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Elasticsearch

...

Elasticsearch is a distributed, RESTful search and analytics engine capable of solving a growing number of use cases.

Use cases

- Full-text search
- Application search
- Enterprise search
- Business analytics
- Metrics analytics
- Security analytics
- Operational logs analytics
- Anomaly detection
- ...

Features overview

- Distributed and scalable
- Resilient
- Fault tolerant
- High availability
- RESTful interface
- Document-oriented
- Schema free
- Multi-tenancy
- Extensible
- Growing and active community

- Query DSL
- Aggregations
- Full-text search (Lucene)
- Structured search
- Geo-spatial search
- Suggesters
- Highlighters
- Percolation
- Profiling
- Client libraries in 10+ languages
- ...

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Terminology (ES vs SQL)

Index ⇔ Database

~~Type ⇔ Table~~

Document ⇔ Record

Field ⇔ Column

Mapping ⇔ Schema

Everything is indexed ⇔ Index

Query DSL ⇔ SQL

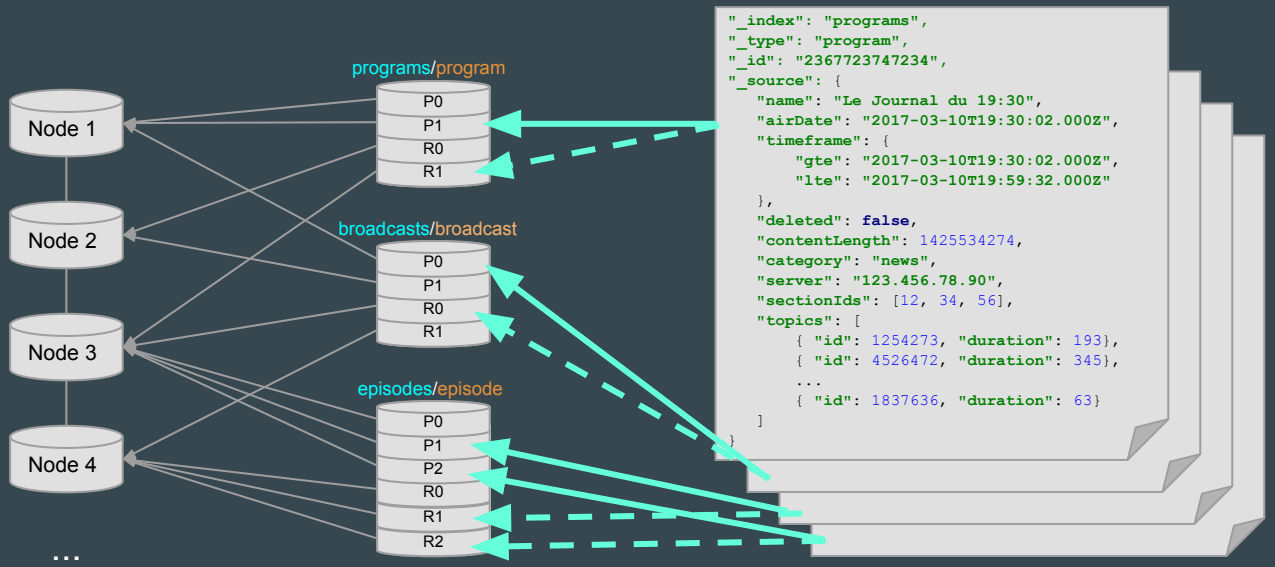
Terminology

Cluster

Indices/Types
& Shards

Documents
& Fields

Schema /
Mapping type



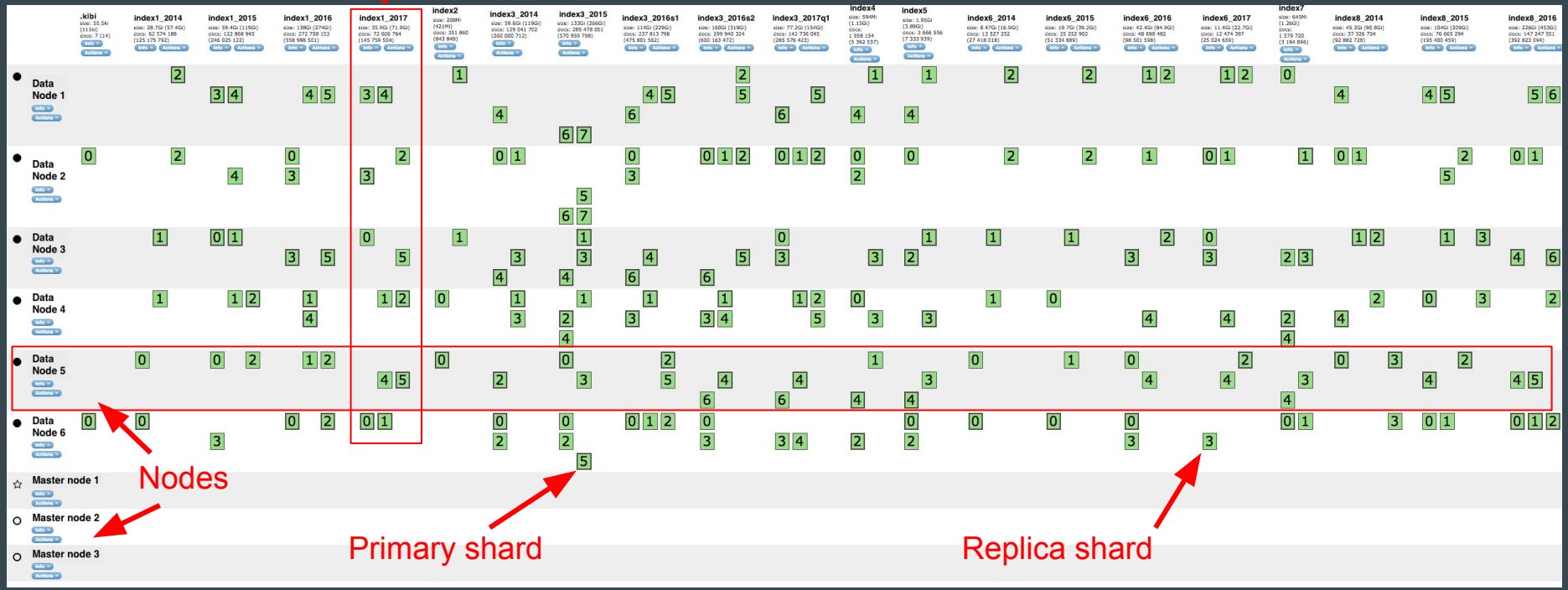

```
"_index": "programs",
"_type": "program",
"_id": "2367723747234",
"_source": {
   "name": "Le Journal du 19:30",
   "airDate": "2017-03-10T19:30:02.000Z",
   "timeframe": {
       "gte": "2017-03-10T19:30:02.000Z",
       "lte": "2017-03-10T19:59:32.000Z"
   },
   "deleted": false,
   "contentLength": 1425534274,
   "category": "news",
   "server": "123.456.78.90",
   "sectionIds": [12, 34, 56],
   "topics": [
       { "id": 1254273, "duration": 193},
       { "id": 4526472, "duration": 345},
       ...
       { "id": 1837636, "duration": 63}
   ]
}
```

```
{
 "program" : {
   "properties" : {
     "airDate" :       { "type" : "date" },
     "category" :      { "type" : "keyword" },
     "contentLength" : { "type" : "long" },
     "deleted" :       { "type" : "boolean" },
     "name" : {
       "type" : "text",
       "fields" : {
         "keyword" :   { "type" : "keyword" }
       }
     },
     "sectionIds" :    { "type" : "long" },
     "server" :        { "type" : "ip" },
     "timeframe" :     { "type" : "date_range" },
     "topics" : {
       "properties" : {
         "duration" :  { "type" : "long" },
         "id" :        { "type" : "long" }
       }
     }
   }
 }
}
```

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Inside a cluster - Overview



Inside a cluster - Different node types

- **Master** node
- **Data** node
- **Ingest** node
- **Coordinating** node
- **Tribe** node (deprecated in v5.3 in favor of cross-cluster search)

Inside a cluster - Master node

- Coordinates the cluster
  - `node.master: true`
  - `node.data:   false`
  - `node.ingest: false`
- Very small footprint (i.e. tiny nodes)
- Quorum: (N / 2) + 1
  - `discovery.zen.minimum_master_nodes: 2`
- 3 master-eligible nodes are usually sufficient

Inside a cluster - Data node

- Workhorse of the cluster
  - `node.master: false`
  - **`node.data: true`**
  - `node.ingest: false`
- Stores the indexed data
- Handles search and aggregation queries
- Performs CRUD operations on indices and documents
- Add more data nodes as your data volume grows => horizontal scaling

Inside a cluster - Ingest node

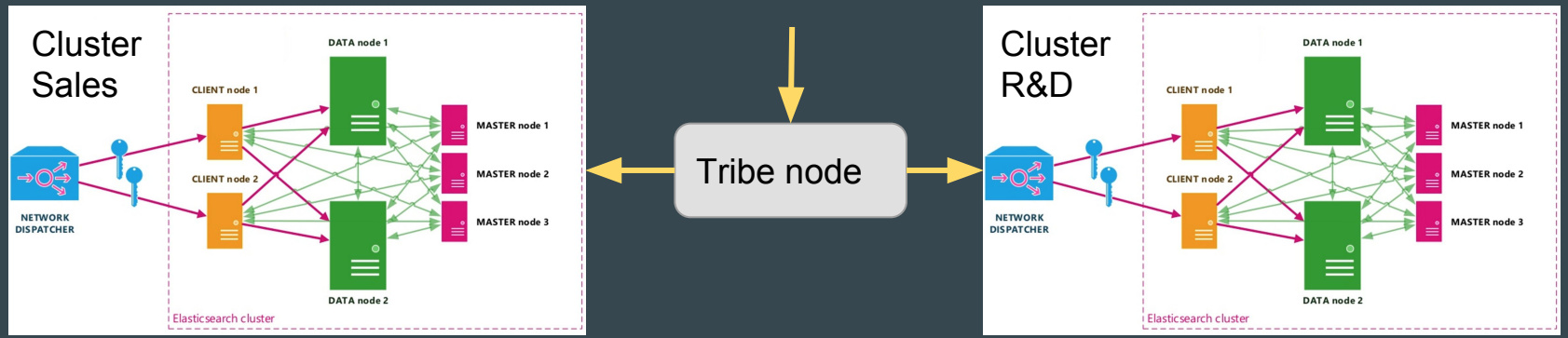
- Pre-processing within the cluster
  - `node.master: false`
  - `node.data: false`
  - **`node.ingest: true`**
- Introduced in ES v5
- Data processing pipeline before the indexing phase
- Supports a large array of processing filters
  - geo, grok, gsub, lowercase, remove, ...
- Similar to a mini/embedded Logstash, yet doesn't replace it

Inside a cluster - Coordinating node

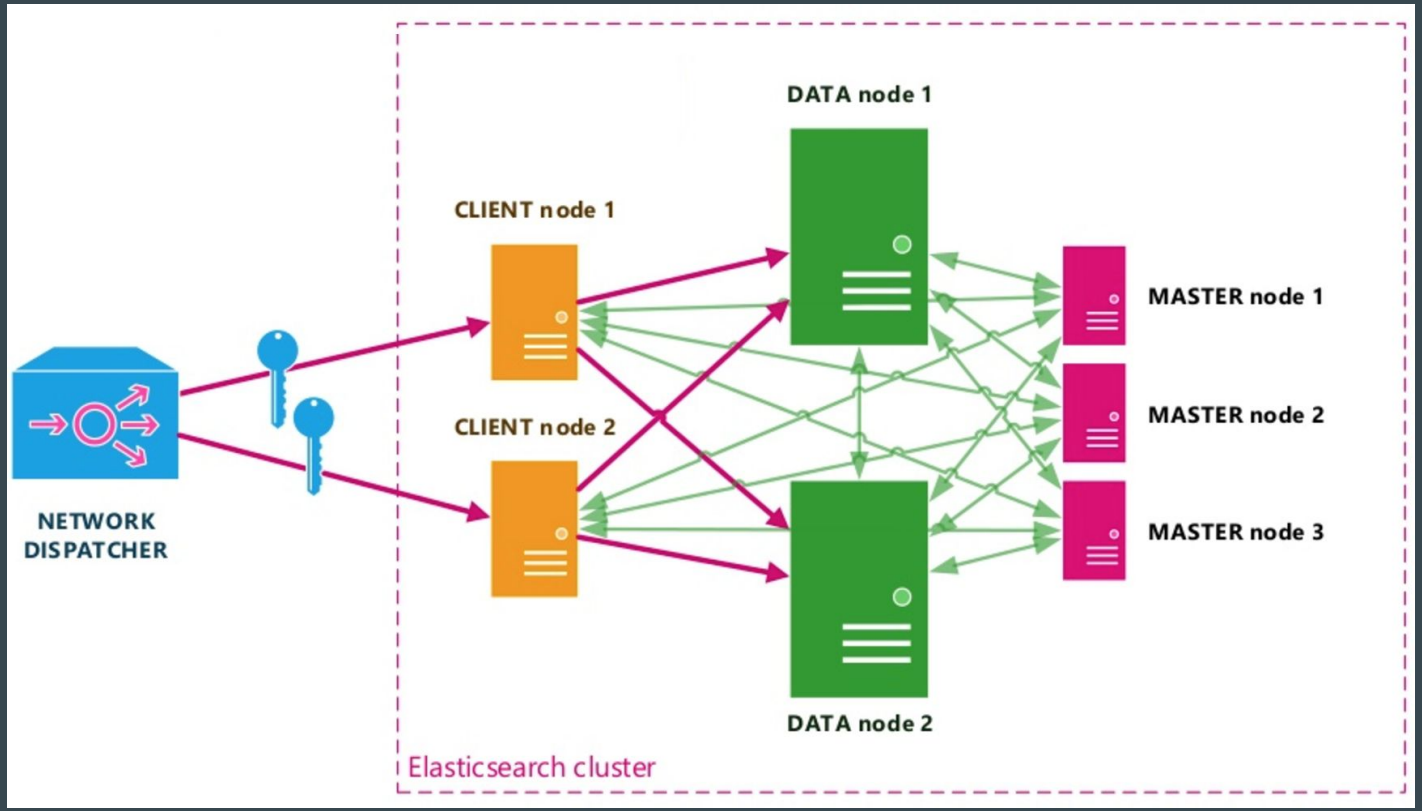
- Load-balancer within the cluster
  - `node.master: false`
  - `node.data:   false`
  - `node.ingest: false`
- aka “client node”
- Routes search requests to data nodes
- Handles the search/reduce phase
- Distributes bulk indexing

Inside a cluster - Tribe node

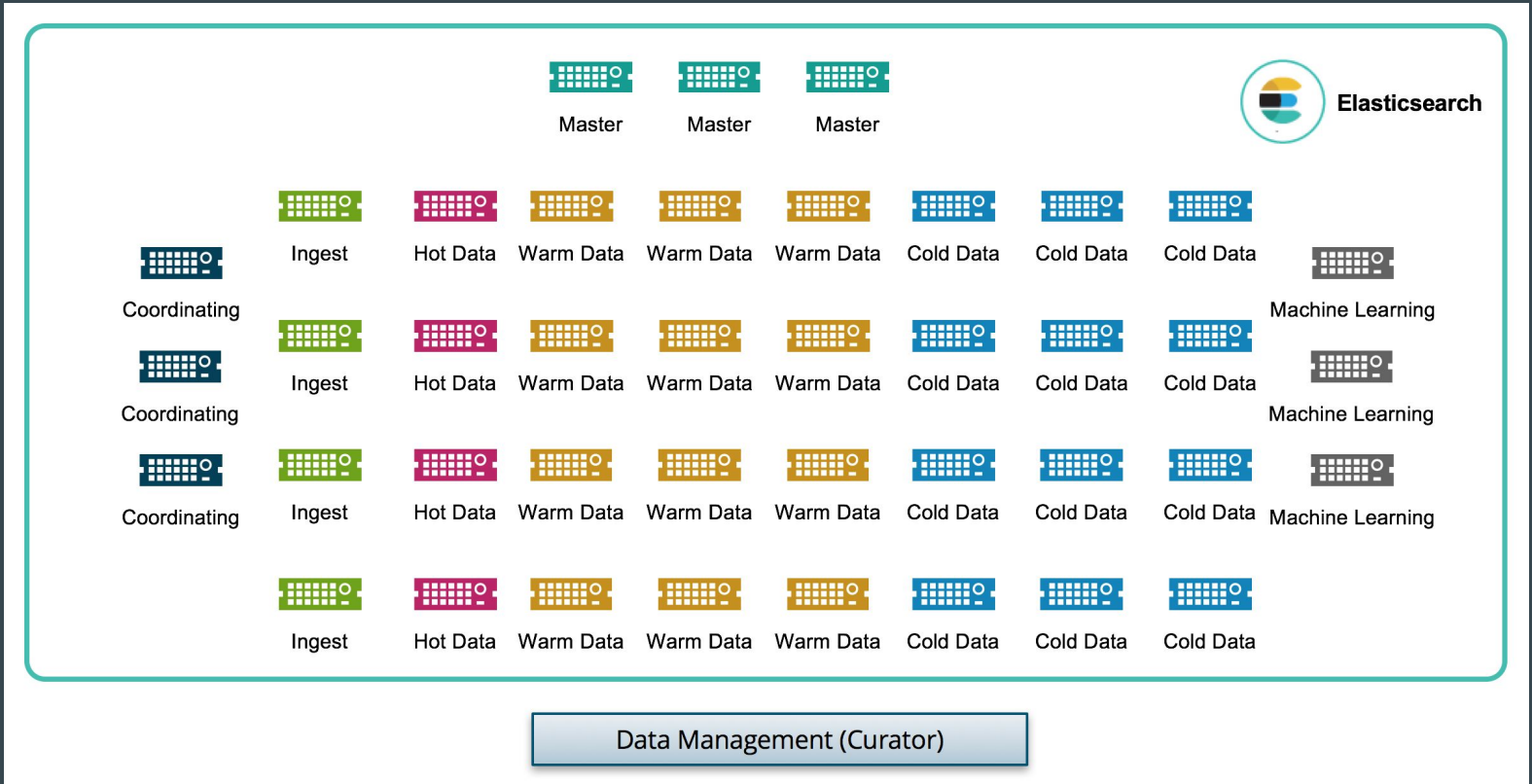
- Federated client between clusters
- Holds global cluster state
- Going away in v5.3 in favor of cross-cluster search
  - https://speakerdeck.com/elastic/whats-evolving-in-elasticsearch-1 (slide 41)



Inside a cluster - Different node types



Inside a cluster - Different node types



Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

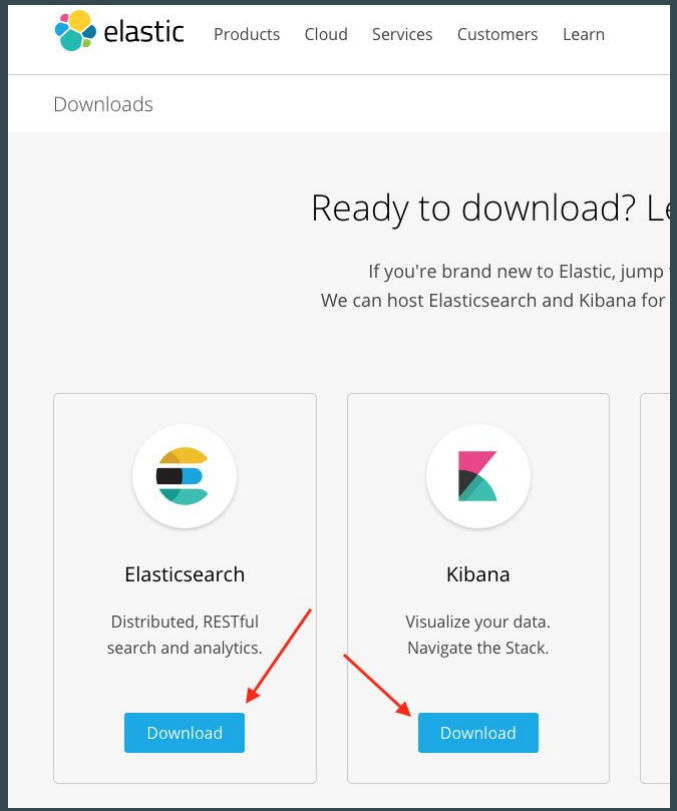
Search & Aggregations

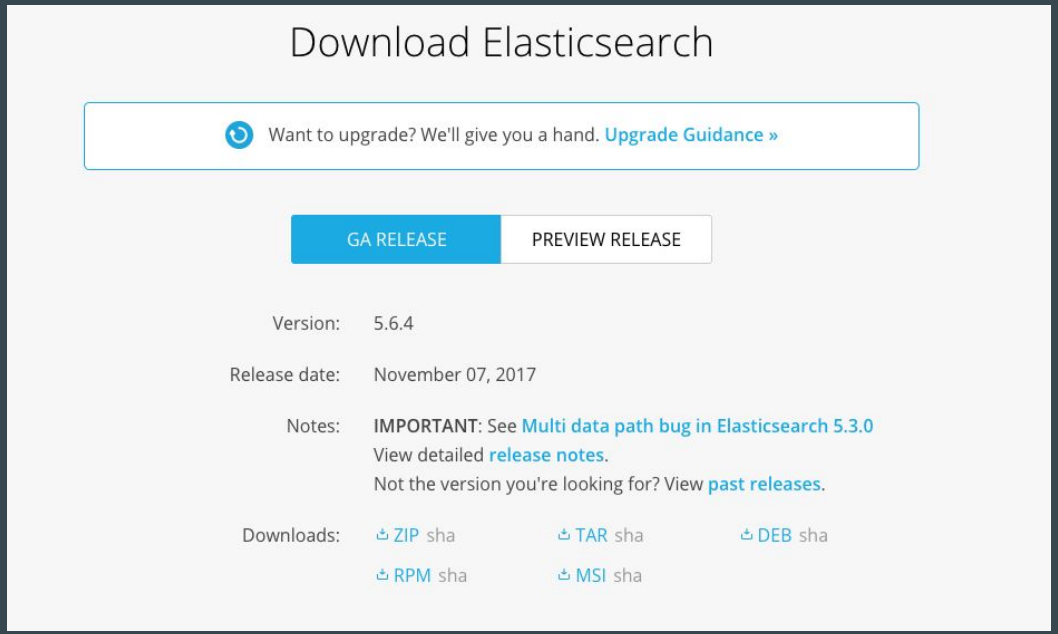
- Structured search
  - `terms, term, range, regexp, wildcard, exists`
  - `bool, must, must_not, filter, should`
  - `geo_shape, geo_distance, geo_bounding_box`
- Full-text search
  - `match, multi_match`
  - `match_phrase, match_phrase_prefix`
  - `query_string`
- Aggregation search
  - Metrics: `sum, avg, stats, min, max, percentile, …`
  - Buckets: `terms, histogram, date_range, ip_range, geo_distance, geohash_grid, …`
  - Pipeline: `moving_avg, serial_diff, sum_bucket, min_bucket, max_bucket, …`
  - Matrix: `matrix_stats`

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Installation - https://elastic.co/downloads





1. Download and unzip
2. ./bin/elasticsearch
3. curl http://localhost:9200/
4. That's all folks !

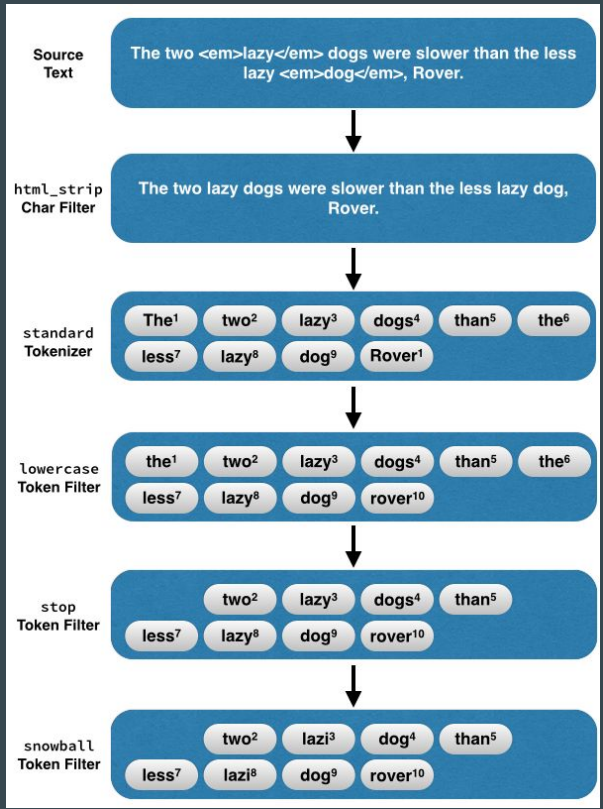
Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Advanced features - Text analysis

Analysis pipeline

- 0-n character filters
  - html_strip, mapping, ...
- 1 tokenizer
  - standard, whitespace, letter, ...
- 0-n token filters
  - lowercase, ngram, unique, elision, ...



Advanced features - Percolation

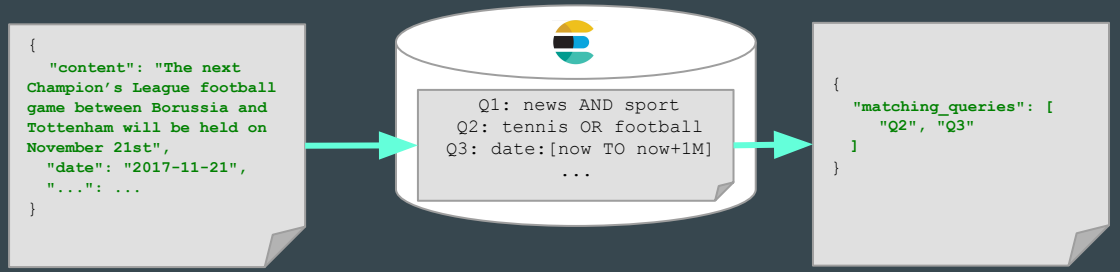
Percolation ⇔ Search in reverse

=> Search: Find the documents that match a given query

=> Percolation: Find the queries that match a given document

Use cases:

- Alerting (pricing, user behavior, ...)
- Classification
- ...



Advanced features - Snapshot/Restore

- Incremental backups
- Multiple type of repositories
  - Filesystem (local / shared)
  - S3
  - HDFS
  - Azure
  - Google Cloud Storage
- Asynchronous
- Great for sourcing new clusters

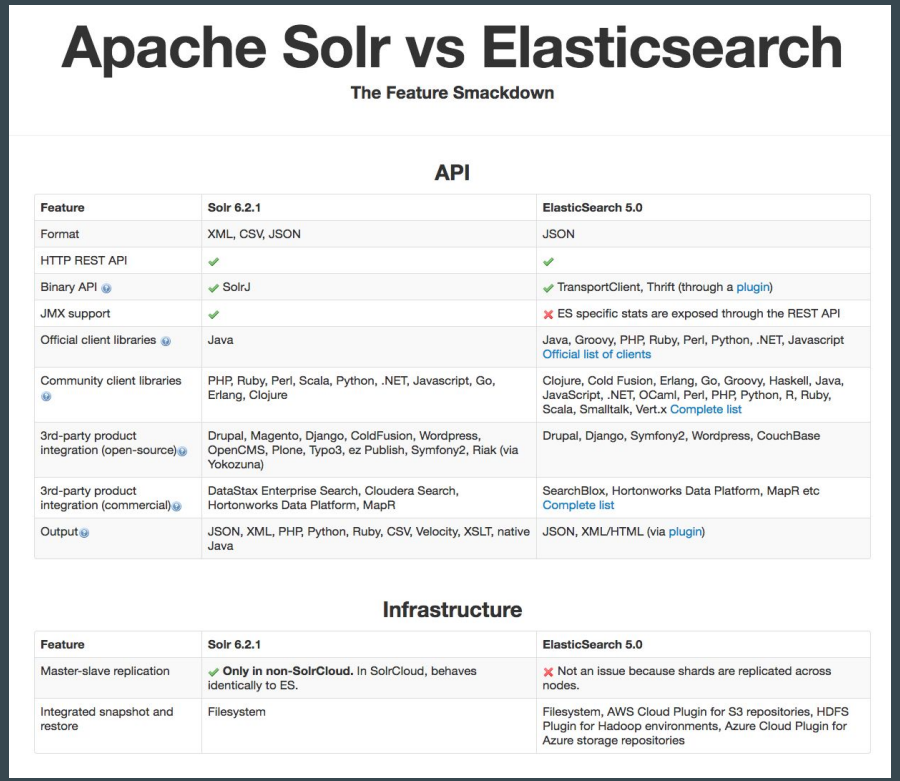
```
$ curl -XPUT 'http://localhost:9200/_snapshot/my_backup' -d '{
    "type": "fs",
    "settings": {
        "location": "/mount/backups/my_backup",
        "compress": true
    }
}'
```

```
PUT /_snapshot/s3_repository?verify=false
{
  "type": "s3",
  "settings": {
    "bucket": "my_s3_bucket",
    "region": "eu-west-1"
  }
}
```

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Differences (1/2)



Apache Solr vs Elasticsearch

The Feature Smackdown

API

Feature	Solr 6.2.1	ElasticSearch 5.0
Format	XML, CSV, JSON	JSON
HTTP REST API	✔	✔
Binary API	✔ SolrJ	✔ TransportClient, Thrift (through a plugin)
JMX support	✔	✖ ES specific stats are exposed through the REST API
Official client libraries	Java	Java, Groovy, PHP, Ruby, Perl, Python, .NET, Javascript Official list of clients
Community client libraries	PHP, Ruby, Perl, Scala, Python, .NET, Javascript, Go, Erlang, Clojure	Clojure, Cold Fusion, Erlang, Go, Groovy, Haskell, Java, JavaScript, .NET, OCaml, Perl, PHP, Python, R, Ruby, Scala, Smalltalk, Vert.x Complete list
3rd-party product integration (open-source)	Drupal, Magento, Django, ColdFusion, Wordpress, OpenCMS, Plone, Typo3, ez Publish, Symfony2, Riak (via Yokozuna)	Drupal, Django, Symfony2, Wordpress, CouchBase
3rd-party product integration (commercial)	DataStax Enterprise Search, Cloudera Search, Hortonworks Data Platform, MapR	SearchBlox, Hortonworks Data Platform, MapR etc Complete list
Output	JSON, XML, PHP, Python, Ruby, CSV, Velocity, XSLT, native Java	JSON, XML/HTML (via plugin)

Infrastructure

Feature	Solr 6.2.1	ElasticSearch 5.0
Master-slave replication	✔ **Only in non-SolrCloud.** In SolrCloud, behaves identically to ES.	✖ Not an issue because shards are replicated across nodes.
Integrated snapshot and restore	Filesystem	Filesystem, AWS Cloud Plugin for S3 repositories, HDFS Plugin for Hadoop environments, Azure Cloud Plugin for Azure storage repositories

http://solr-vs-elasticsearch.com/

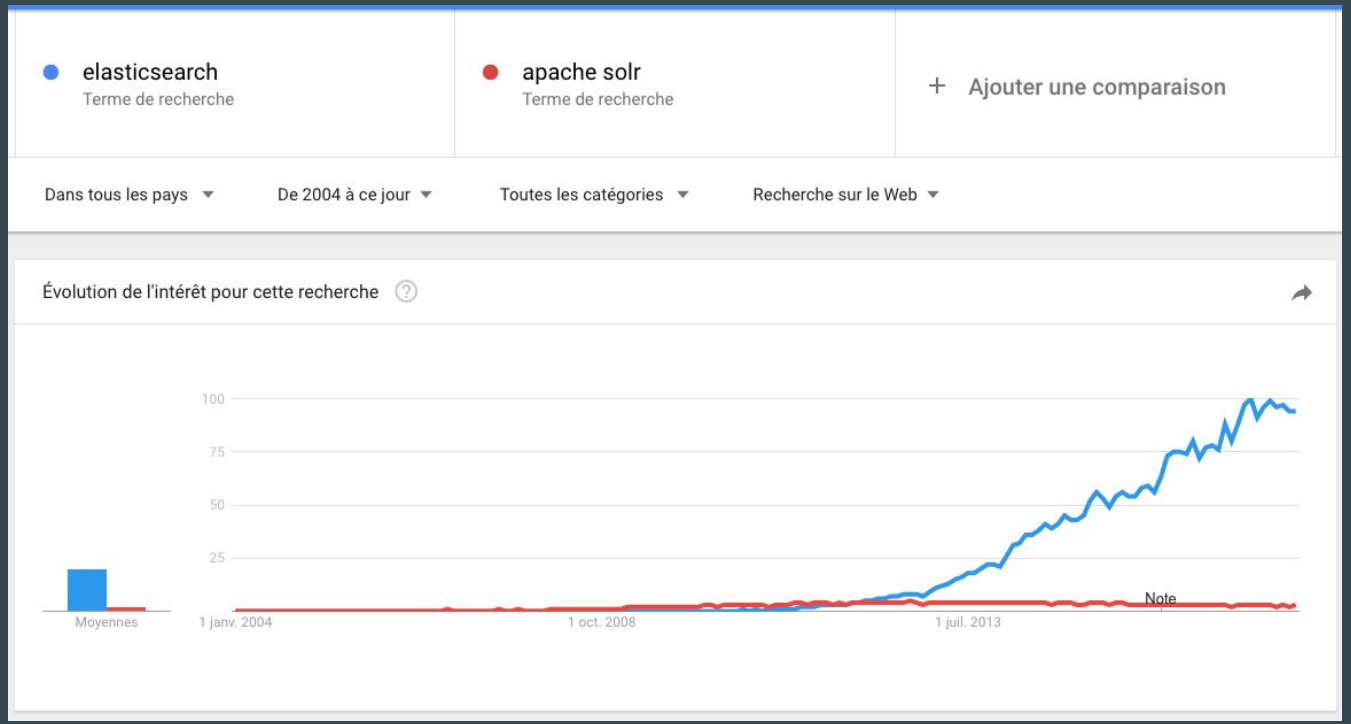
Differences (2/2)

Features	Solr	Elasticsearch
Query Language	Lucene only	Lucene + DSL
Inter-index joins	Yes	No
Percolation	No	Yes
Cluster	Needs Zookeeper	Self-contained
Shard rebalancing	Manual	Automatic
Shard number	Can be increased	Cannot be increased
Ecosystem	Limited	Rich
Snapshot / Restore	Filesystem only	FS + Azure + S3 + GCS + ...

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

Trends (1/4)



https://trends.google.ch/trends/explore?date=all&q=elasticsearch,apache%20solr

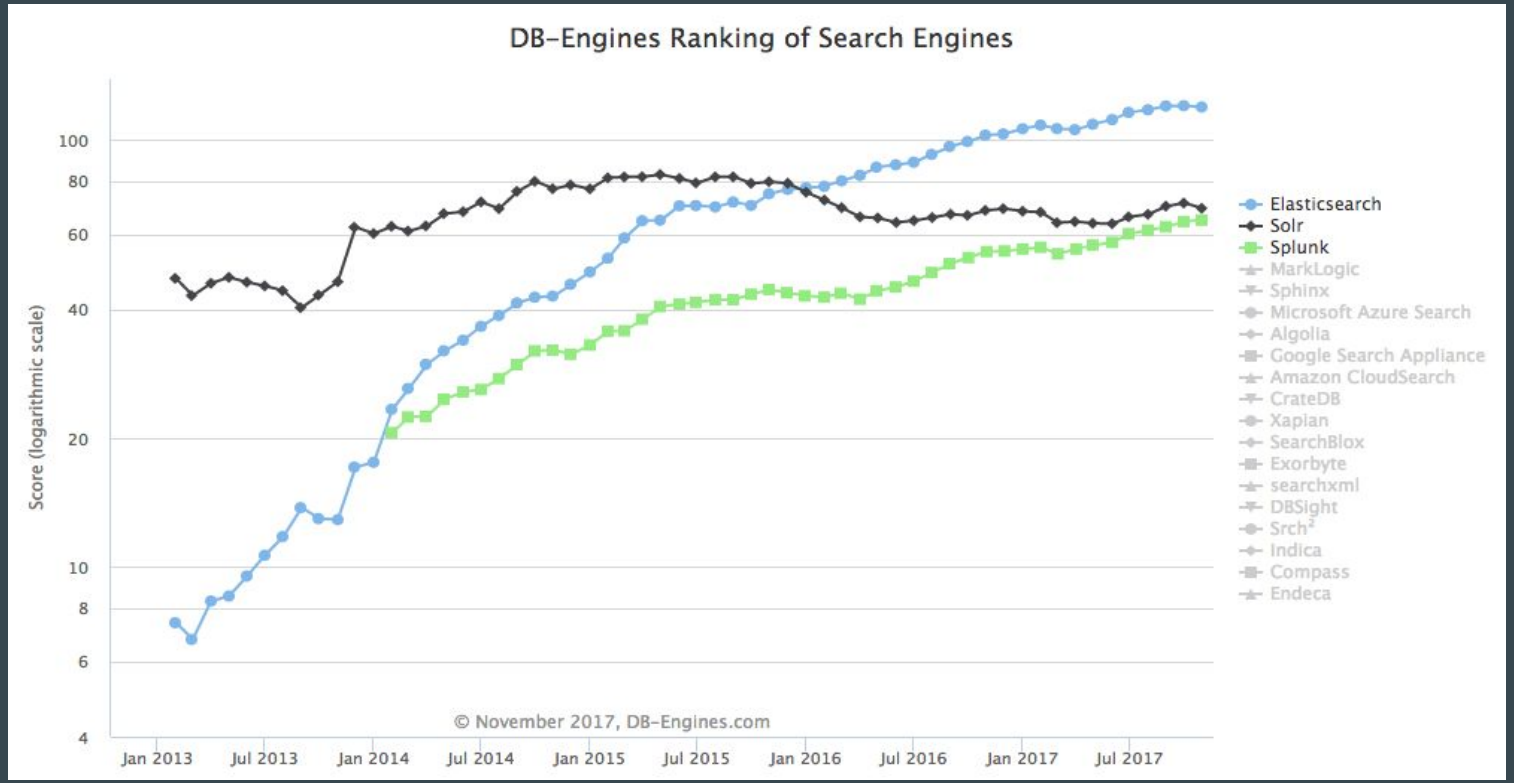
Trends (2/4)

16 systems in ranking, November 2017

Rank			DBMS	Database Model	Score		
Nov 2017	Oct 2017	Nov 2016			Nov 2017	Oct 2017	Nov 2016
1.	1.	1.	Elasticsearch	Search engine	119.41	-0.82	+16.84
2.	2.	2.	Solr	Search engine	69.16	-1.97	+0.80
3.	3.	3.	Splunk	Search engine	64.87	+0.51	+10.14
4.	4.	4.	MarkLogic	Multi-model	11.55	-0.26	+1.33
5.	5.	5.	Sphinx	Search engine	5.88	-0.14	-1.11
6.	6.	↑ 8.	Microsoft Azure Search	Search engine	3.88	+0.20	+1.96
7.	↑ 8.	↑ 9.	Algolia	Search engine	2.78	+0.14	+1.23
8.	↓ 7.	↓ 6.	Google Search Appliance	Search engine	2.76	-0.06	+0.14
9.	9.	↓ 7.	Amazon CloudSearch	Search engine	2.36	-0.04	+0.10
10.	10.	↑ 12.	CrateDB	Multi-model	0.66	-0.08	+0.43
11.	11.	↓ 10.	Xapian	Search engine	0.57	-0.02	+0.06
12.	12.	↑ 14.	SearchBlox	Search engine	0.24	-0.01	+0.13
13.	13.	↑ 16.	Exorbyte	Search engine	0.06	-0.02	+0.06
14.	↑ 15.	↑ 16.	searchxml	Multi-model	0.04	+0.04	+0.04
15.	↓ 14.	15.	DBSight	Search engine	0.03	-0.01	+0.02
16.	↓ 15.	↓ 11.	Indica	Search engine	0.00	±0.00	-0.36

https://db-engines.com/en/ranking/search+engine

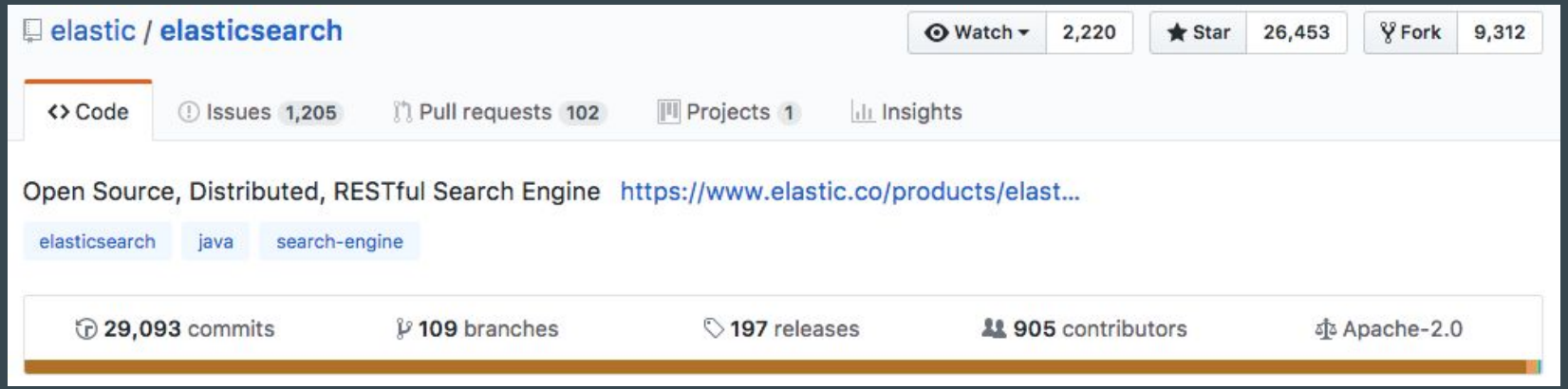
Trends (3/4)



https://db-engines.com/en/ranking_trend/search+engine

Trends (4/4)


apache / lucene-solr
mirrored from git://git.apache.org/lucene-solr.git
Watch 195
Star 1,234
Fork 1,099
Code
Pull requests 143
Projects 0
Insights
Mirror of Apache Lucene + Solr
28,966 commits
83 branches
237 releases
70 contributors


elastic / elasticsearch
Watch 2,220
Star 26,453
Fork 9,312
Code
Issues 1,205
Pull requests 102
Projects 1
Insights
Open Source, Distributed, RESTful Search Engine https://www.elastic.co/products/elast...
elasticsearch
java
search-engine
29,093 commits
109 branches
197 releases
905 contributors
Apache-2.0

Agenda

- Getting acquainted
- Ecosystem
- Features overview
- Terminology
- Inside a cluster
- Search & aggregations
- Installation
- Advanced features
- Differences
- Trends
- What's next?

What's next?

- Docs: https://www.elastic.co/guide/index.html
- Videos: https://www.elastic.co/videos
- Slides: https://speakerdeck.com/elastic/
- Blog: https://www.elastic.co/blog
- Source: https://github.com/elastic
- Meetups: https://www.meetup.com/fr-FR/elastic-switzerland/
- Conference: https://www.elastic.co/elasticon/conf/2018/sf
- Discuss:
  - http://stackoverflow.com/questions/tagged/elasticsearch
  - https://discuss.elastic.co/
- Comparisons:
  - http://solr-vs-elasticsearch.com/
  - https://logz.io/blog/solr-vs-elasticsearch/
  - https://sematext.com/blog/solr-vs-elasticsearch-differences/
  - https://sematext.com/blog/solr-elasticsearch-comparison/

Q&A