

Liferay Boot Camp 2019



Elastic & Liferay: come potenziare l'integrazione
con potenzialità di search avanzato



Stefano Pampaloni



Nicola Pagni

Agenda

- Elastic Stack Overview
- Text Analysis and Mappings
- Advanced Search

Elastic Stack Overview

What is Elasticsearch ?

Elastic offers a wide
set of **solutions**



Based on the
leading search engine



Complemented by advanced
security & analytics features



5,500

Customers

And 225M+ downloads

+1k

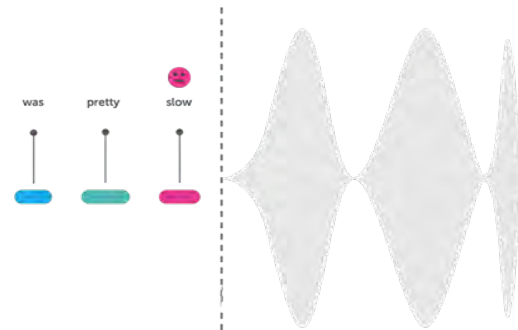
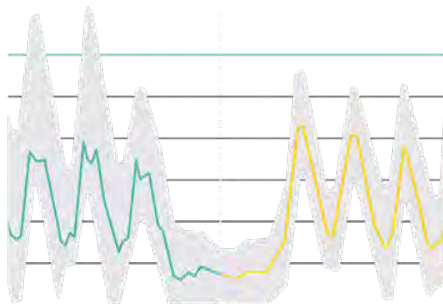
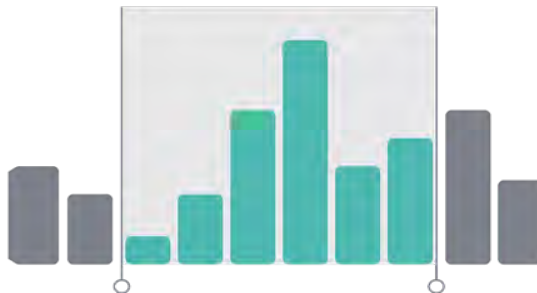
People WW

Inc. 400 @engineering

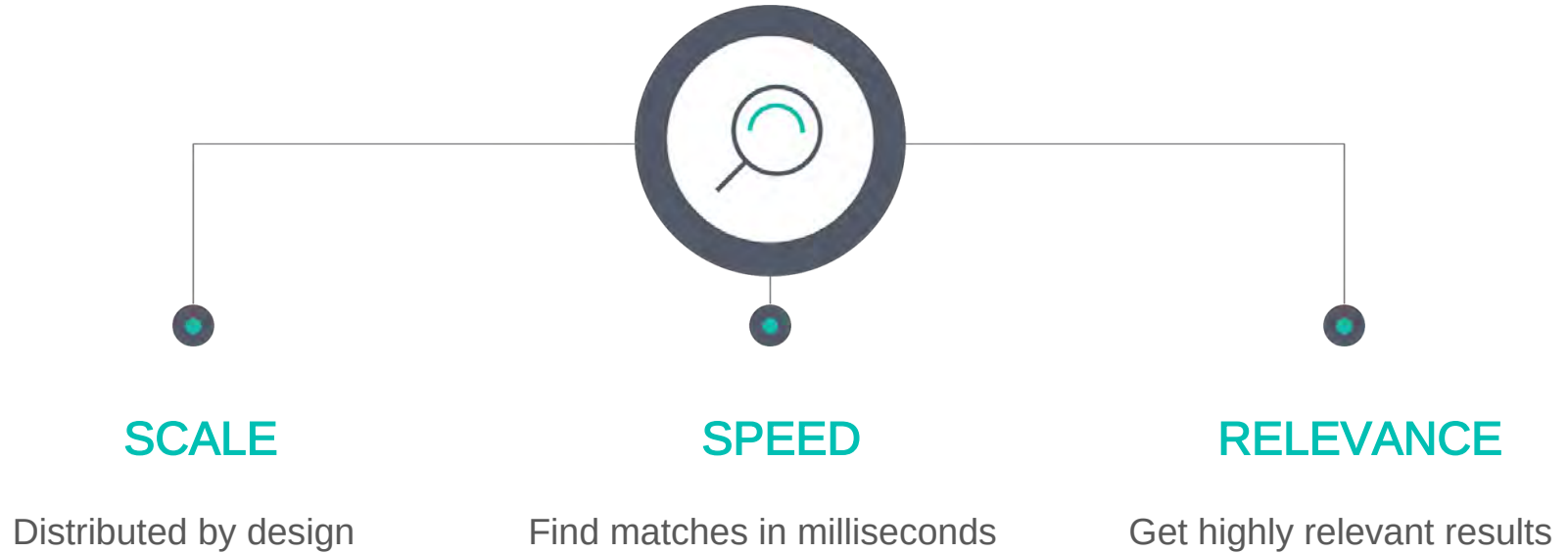
Search is a constant/foundation



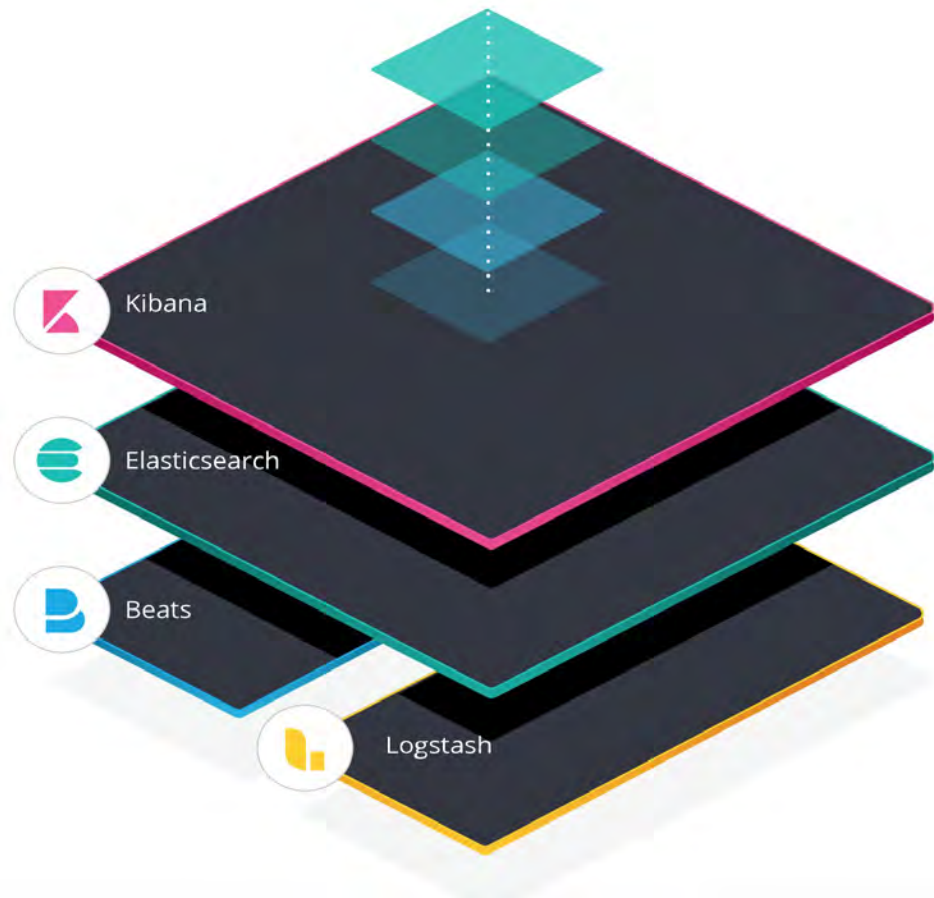
.54 seconds | 1,000,000,000 records



Technology differentiation



Elastic Stack



Solutions

Logging	Metrics	APM	Security Analytics	FUTURE
App Search	Site Search	Enterprise Search	Business Analytics	



Kibana



Visualize &
Manage

Elasticsearch



Store, Search, &
Analyze

Beats



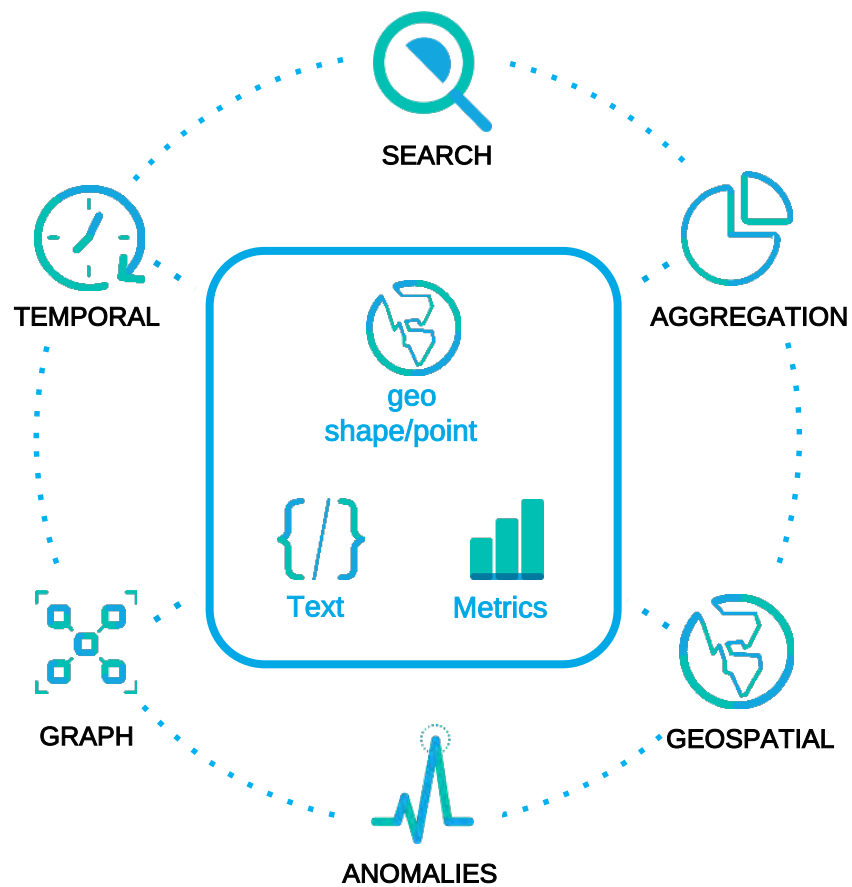
Logstash

Ingest



Elasticsearch

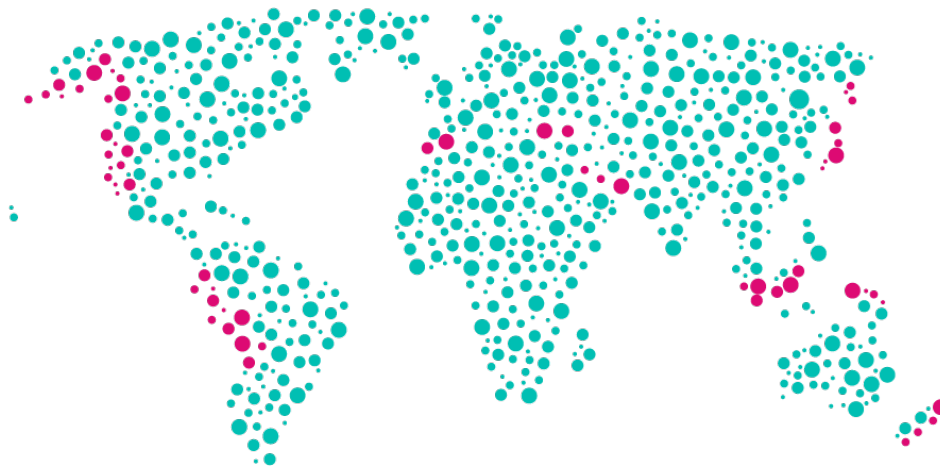






Kibana

Window into the Elastic Stack



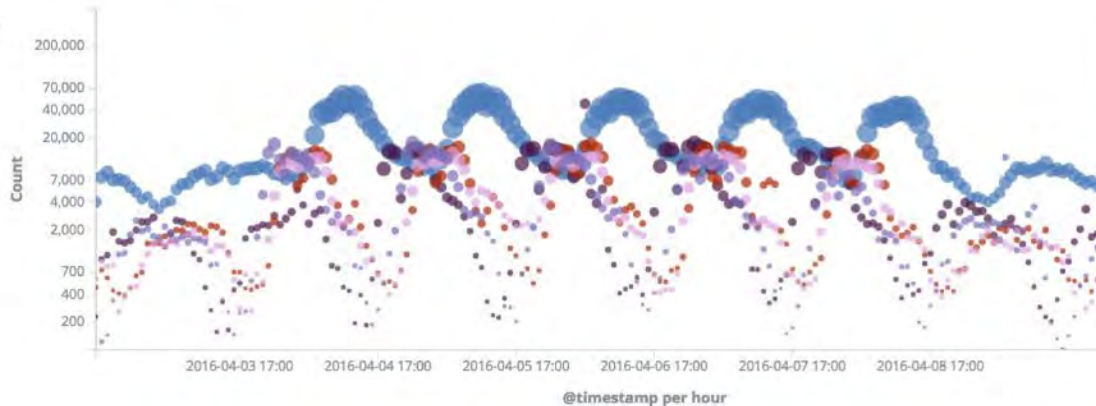
Apache - Bytes and Count

Count

Apache - Unique Visitors

128,088

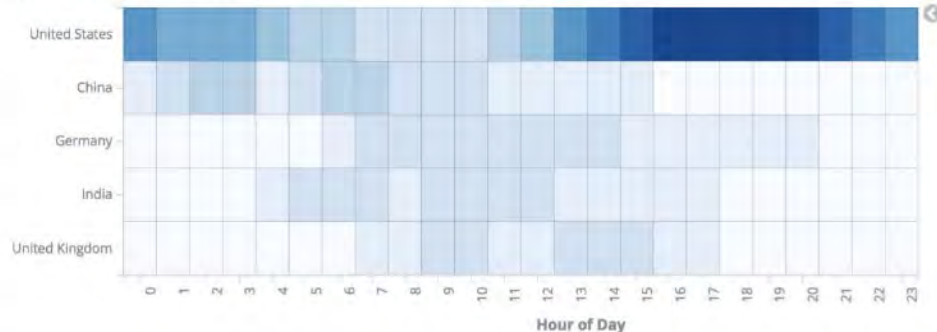
Unique Visitors



Apache - Unique Visitors by City

City 🏙️	Unique Visitors 📈
Beijing	243,706
Woodbridge	134,791
Seattle	123,545
Seoul	111,663
Mumbai	109,801
Tokyo	106,182
New York	102,947
Bangalore	102,738

Apache - Country traffic by hour



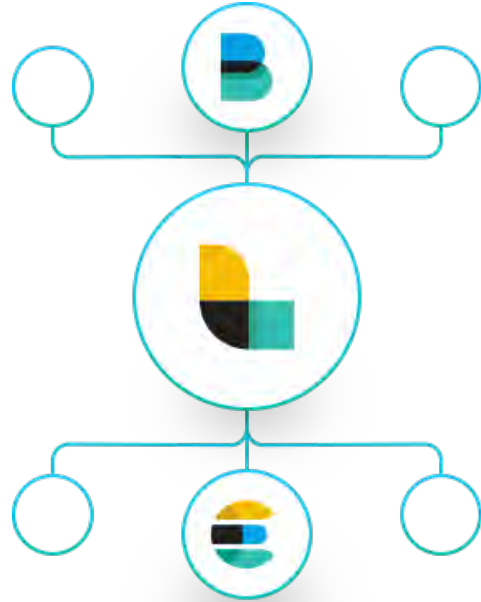
Apache - Top OS





Logstash

Data processing pipeline



ETL

Extract

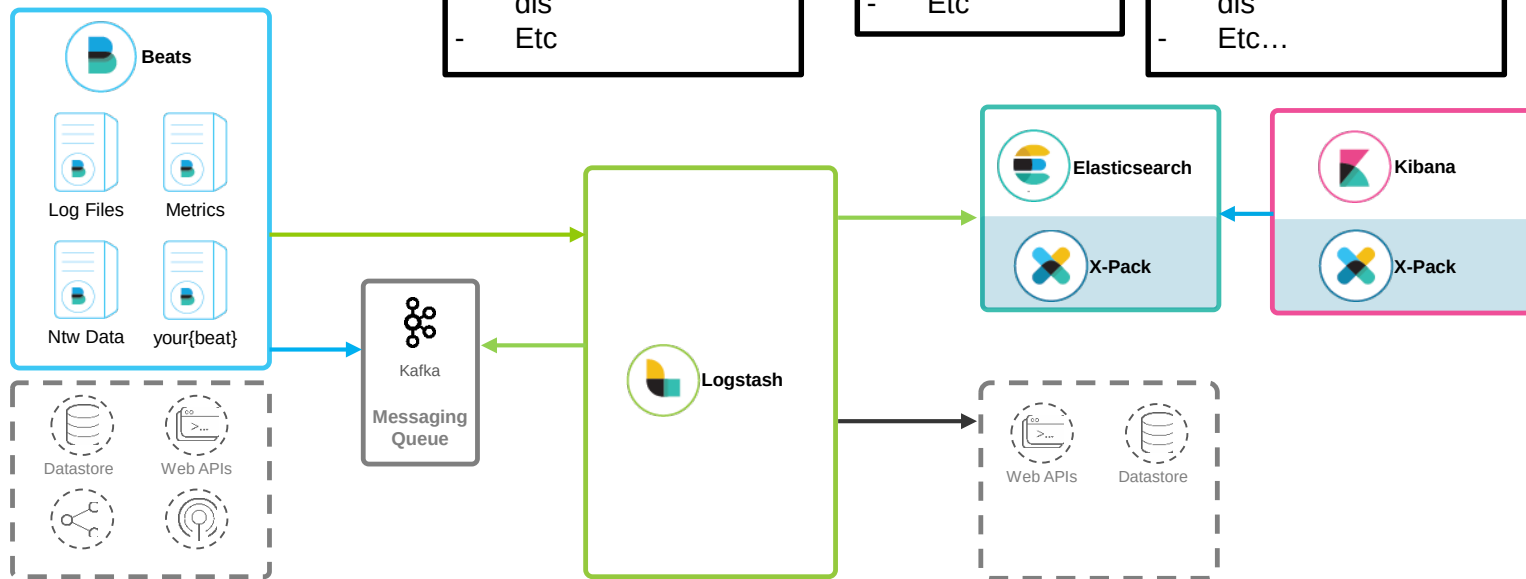
- Beats
- Syslog
- File
- Graphite
- API
- JDBC (SQL)
- Kafka/RabbitMQ/Redis
- Etc

Transform

- Parse
- Extract
- Decode
- Aggregate
- Mutate
- Enrich
- Drop
- Etc

Load

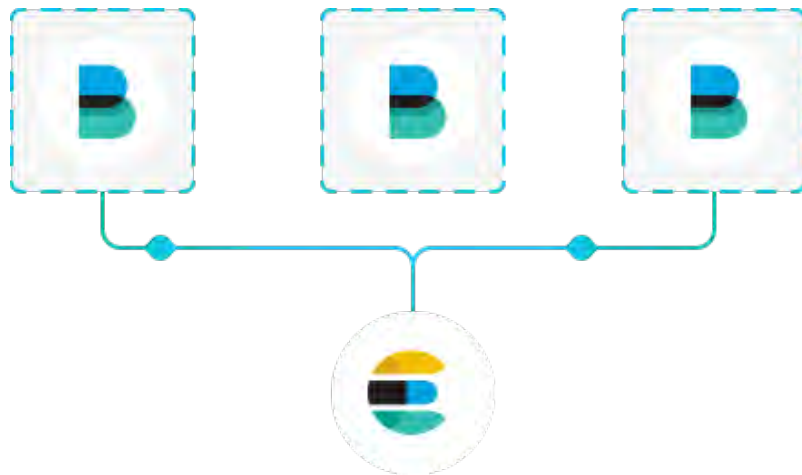
- Elasticsearch
- Syslog
- File/CSV/S3/stdout
- Email
- API
- JDBC (SQL)
- Kafka/RabbitMQ/Redis
- Etc...





Beats

Lightweight data shippers





FILEBEAT
Log Files



METRICBEAT
Metrics



PACKETBEAT
Network Data



WINLOGBEAT
Window Events



AUDITBEAT
Audit Data



HEARTBEAT
Uptime Monitoring

Plus, more than 30 community Beats and growing ...



beat.hostname:"mar.local"

Uses lucene query syntax

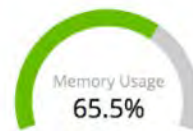
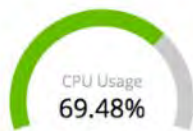
Add a filter +

System Navigation

System Overview | Host Overview (requires Kibana > 5.4) | Docker containers overview

Warning

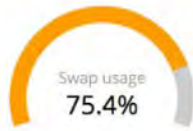
This dashboard is only compatible with Kibana 5.4 and above.



Inbound Traffic
1.5KB/s
Total Transferred 2.3MB

Outbound Traffic
1.4KB/s
Total Transferred 1.1MB

In Packetloss
0
Out Packetloss 3

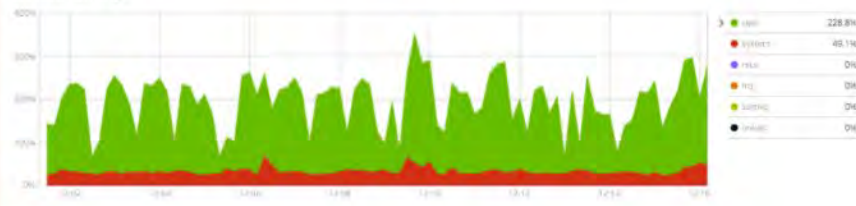


Memory usage
10.5GB
Total Memory 16.0GB

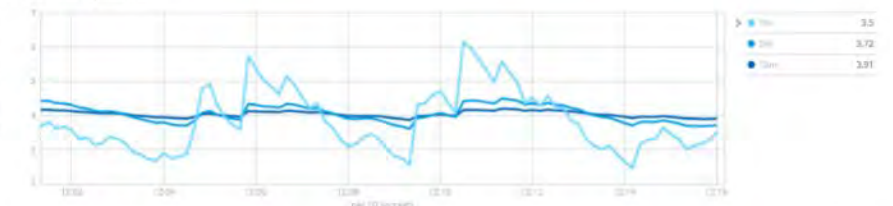
Number of processes
Processes
19
Processes



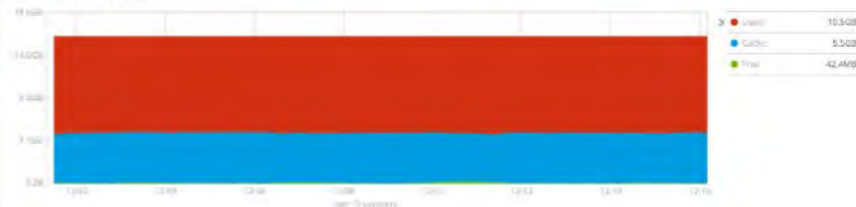
Metricbeat: CPU Usage



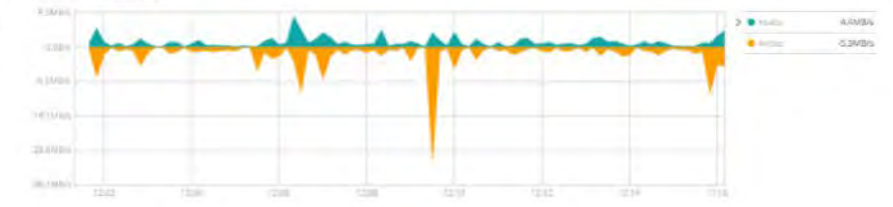
Metricbeat: System Load



Metricbeat: Memory Usage



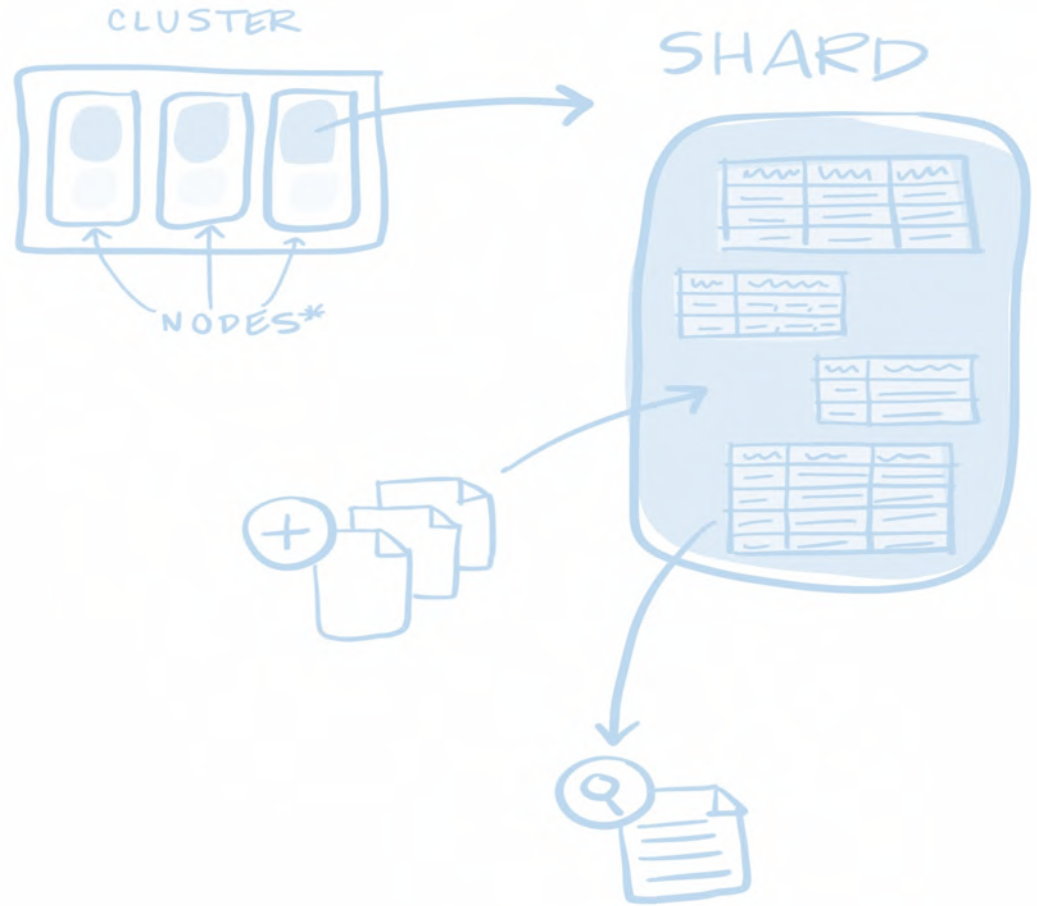
Metricbeat: Disk IO (Bytes)



Text Analysis and Mappings

Basic Concepts

- Cluster
- Nodes
- Indexes
- Mappings
- Type



Mappings Datatype

Basic

- text
- date
- boolean
- integer
- float
- ...

Advanced

- keyword
- geo_point
- ip
- range
- nested
- join

Automatic Mappings Creation

- The index operation automatically creates an index if it has not been created before.
- Elasticsearch try to guess datatypes for each fields.
- Strings are mapped twice: text & keyword (these fields are named *Multi-fields*).

```
{
  "lrbc19": {
    "mappings": {
      "doc": {
        "properties": {
          ...
          "speech_name": {
            "type": "text",
            "fields": {
              "keyword": {
                "type": "keyword",
                "ignore_above": 256
              }
            }
          },
          ...
        }
      }
    }
  }
}
```

Text Analysis

- Strings need to be analyzed, sometimes not. It depends on how we want to use them.
 - ❑ *Full-text search / Exact Search*
 - ❑ *Analytics*
 - ❑ *Sorting*
- There are several way to analyze a string, through analyzers:
 - ❑ *Standard / Whitespace / Keyword / Simple/ Language etc...*

Analyzer

- How analysis works?

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date:
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimmed,
And every fair from fair sometime declines,
By chance, or nature's changing course untrimmed:
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date:
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimmed,
And every fair from fair sometime declines,
By chance, or nature's changing course untrimmed:
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.

Term	Doc 1	Doc 2	Doc 3
breathe			
brings			
buds			
but			
by			
can			
...			
damasked			
darling			
date			
day			
deaf			
death			
declines			
delight			

sorted list of
unique terms

Custom Analyzer



Token Filters

- The most popular is *lowercase*.
- There are many built-in token filters, such as :
 - ❑ Stopwords
 - ❑ Elision
 - ❑ Stemming
 - ❑ Synonyms
 - ❑ ...and customs!

Token Filters Definitions

Zero coding!

Stopwords

```
{  
  "my_stop": {  
    "type": "stop",  
    "stopwords": ["and", "is", "the"]  
  }  
}
```

Elision

```
{  
  "elision" : {  
    "type" : "elision",  
    "articles" : ["I", "m", "t", "qu", "n", "s", "j"]  
  }  
}
```

Stemming

```
{  
  "my_stemmer" : {  
    "type" : "stemmer",  
    "name" : "german"  
  }  
}
```

Synonyms

```
{  
  "synonym" : {  
    "type" : "synonym",  
    "synonyms_path" : "analysis/synonym.txt"  
  }  
}
```

Token Filters by examples

- **Elision**

1

Define a specific mapping for the field “description”.

2

Index a documents { description:“Nicola’s laptop is nice” }

3

Typing “nicola” or “nicola’s” we have same results.

- **Stemming**

2

Index a documents { description:“The bootcamp participants are excellent developers” }

3

Typing “developer” or “developers” we have same results.

Token Filters by examples

- **Synonyms**

1

Define a specific mapping, creating a synonyms file and save it in each node.

❏ Example file of Synonyms:

painting, picture, tableau...

2

Index a documents { description: "This painting is really beautiful." }

3

Typing "painting" or "picture" we have same results.

Tokenizer by examples

- **Edge-Ngram**

- ❑ It breaks text down into words whenever it encounters one of a list of specified characters

- ❑ It emits N-grams of each word where the start of the N-gram is anchored to the beginning of the word.

...perfect for search-as-you-type feature!!

Advanced Search

QueryDSL

Basic

- Match
 - ❑ match_all / multi_match / match_phrase
- Terms
- Range

Advanced

- Geo Distance
- Boolean
- Proximity
- Fuzziness
- MLT
- Common

QueryDSL - Match

- The Match query allows us to do free text searches - *through the analysis defined*

```
GET index_name/_search
{
  "query": {
    "match": {
      "description": "dEVEloperS"
    }
  }
}
```


QueryDSL - Match & Fuzziness

- It allows us to still have results despite the errors

```
GET index_name/_search
{
  "query": {
    "match": {
      "field_name": {
        "query": "D3velopers",
        "fuzziness": 1
      }
    }
  }
}
```

```
GET index_name/_search
{
  "query": {
    "match": {
      "field_name": {
        "query": "D3veloprs",
        "fuzziness": 2
      }
    }
  }
}
```

QueryDSL - Match & Proximity

- It allows us to have results based on the distance - *slop* - of the tokens

The XXV Acea Rome International Marathon, the running event organized by the Italian Athletics Federation and by Roma Capitale, started in the rain. The marathon (opened with the disabled race) started from Via dei Fori Imperiali, at Largo Corrado Ricci, and arrived in Via dei Fori Imperiali just before the intersection with Via Cavour

...

```
GET index_name/_search
{
  "query": {
    "match": {
      "field_name": {
        "query": "MARathon International",
        "slop": 2
      }
    }
  }
}
```

QueryDSL - Rank Feature



- It allows us to **boost the score of documents** based on the values of numeric features.
- They work on **rank_feature** datatype.

mapping

```
"properties": {  
  "pagerank": {  
    "type": "rank_features"  
  },  
  ...  
}
```

indexing

```
PUT lrbc19/_doc/1  
{  
  "url": "www.seacom.it",  
  "content": "some content here...",  
  "pagerank": 50.3  
}
```

query

```
GET lrbc19/_search  
{  
  "query": {  
    "bool": {  
      "must": [...],  
      "should": [  
        {  
          "rank_feature": {  
            "field": "pagerank"  
          }  
        }  
      ]  
    }  
  }  
}
```

QueryDSL - Multi Match & Boosting

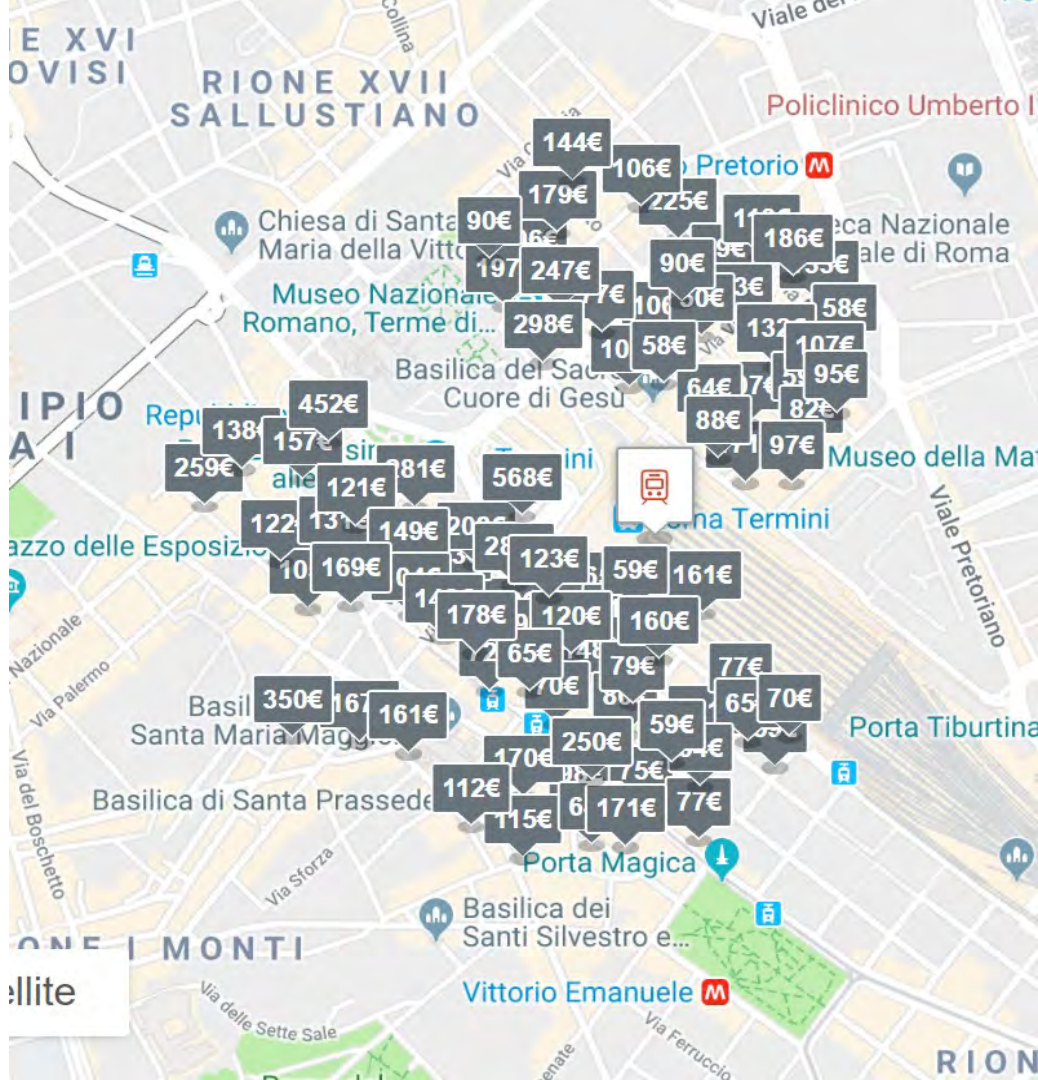
- It allows us to do free text search on more than one field

```
GET index_name/_search
{
  "query": {
    "multi_match": {
      "query": "marathon ROMA",
      "type": "best_fields",
      "fields": [
        "title^2",
        "content"
      ]
    }
  }
}
```

QueryDSL - Geo Distance

```
GET /index name/ search
```

```
{
  "query": {
    "bool": {
      "must": {
        "match_all": {}
      },
      "filter": {
        "geo_distance": {
          "distance": "3km",
          "pin.location": [41.9, 2.5]
        }
      }
    }
  }
}
```



QueryDSL - MLT

- The **MLT** query (*More Like This*) finds documents that are “like” a given set of documents or pieces of text

```
GET /_search
{
  "query": {
    "more_like_this": {
      "fields": ["title", "description"],
      "like": [
        {
          "_index": "lrbc19",
          "_id": "1"
        }
      ]
    }
  }
}
```

Domande



CONTATTI

Stefano Pampaloni

Email stefanop@seacom.it

Skype anteosnet



Nicola Pagni

Email nicolap@seacom.it

Skype nicola.pagni.seacom



Liferay Boot Camp 2019

EATALY | ROMA
16 MAGGIO

Grazie! 🙏