Introduction to

elasticsearch

By Melvyn Peignon

What will I cover?

- Company and products presentation
- Elasticsearch architecture
- Presentation of Kibana
- Presentation of the search API
- Analyzer
- TF/IDF and relevance
- Elasticsearch use case
- Conclusion

Elastic

Founded in 2012

- Is behind:
  - Kibana
  - Elasticsearch
  - Logstash
  - Beats

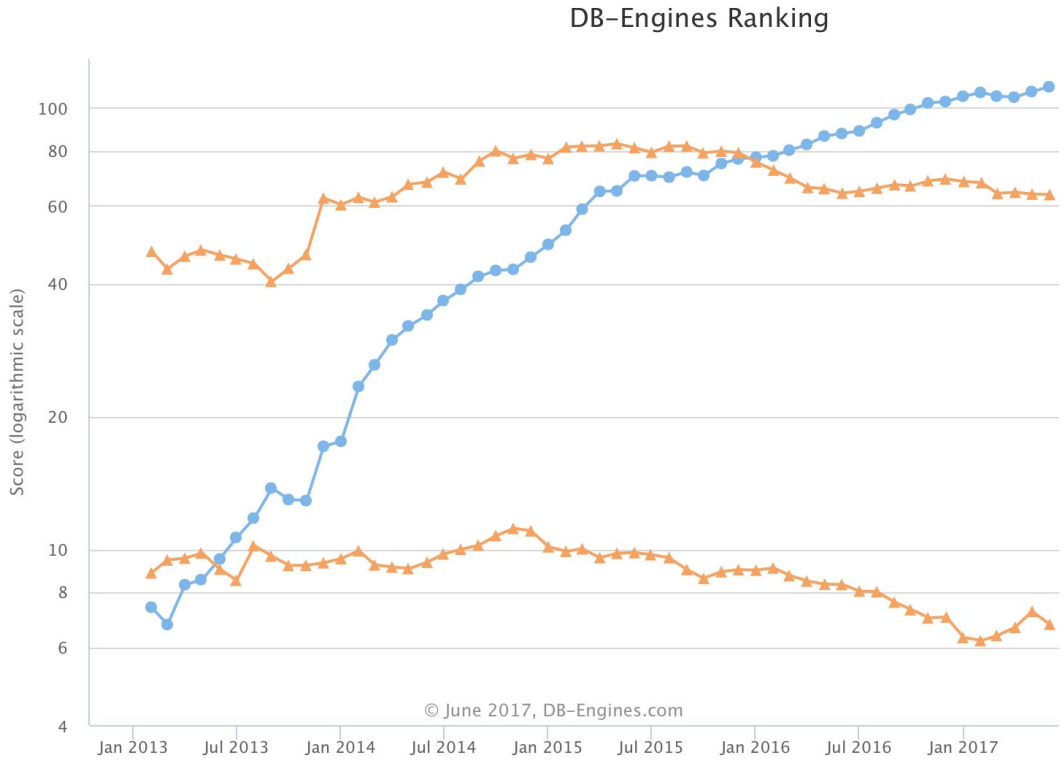
logstash

What is elasticsearch?

- Full text search engine
- Based on Lucene
- Highly available
- Distributed
- Scalable
- RESTful
- Open Source

Shay
Bannon

Trending between search-engine (ES is blue)



How do they make money?

Objectives

This instructor-led in-classroom course will start with an introduction to Elasticsearch and the Search API. You will also learn how to use text analysis, define mappings, understand the distributed model of Elasticsearch, and how to manipulate search results. This course also covers aggregations in detail and how to handle data relationships. Through a series of extensive hands-on labs, you will ingest data into Elasticsearch, write search queries, define mappings, and perform aggregations as you work your way through the various skills needed of an Elasticsearch developer.

Core Elasticsearch:

Sep 4 HKT in HONG KONG, China

To register for this class, please provide participant information below. Registrations will close 2 days prior to the start of the course.

Registrations placed before Sunday, July 16, 2017 11:59 PM HKT are priced at $ 1360.00 USD.

The regular price for this training is $ 1600.00 USD per participant.

By submitting this form you are confirming that you have read and agree to our Privacy Policy.

Register

☐ This is for someone else

CRUD

CREATE

READ

UPDATE

DELETE

Some concepts to know

- Near real time (NRT)
- Cluster
- Node
- Index
- Document
- Shards and Replicas

Documents, Types, indexes

- An index is a collection of documents that share similar properties.
- A document is the basic piece of information that can be indexed.
- A type is a logical partition of the data in your index

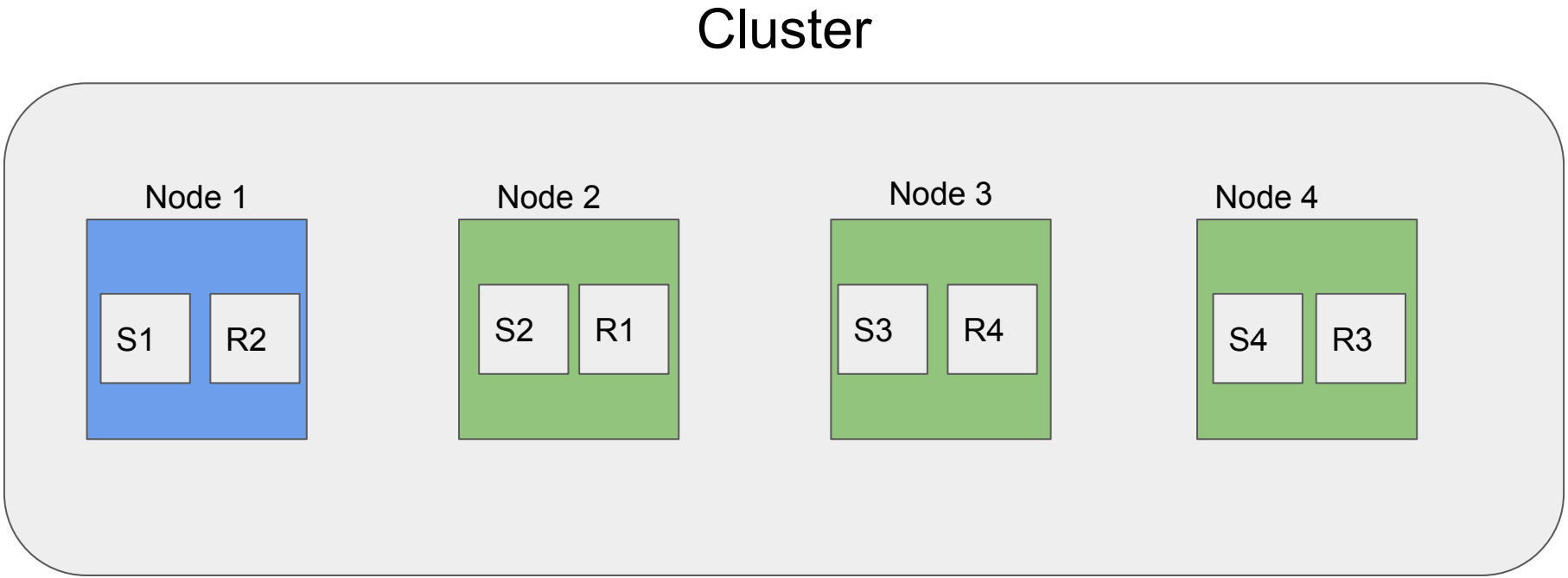
```
{
  "name": "Bill",
  "age": 20,
  "profession": "Architect"
}
```

Cluster, Nodes, Shards and Replicas

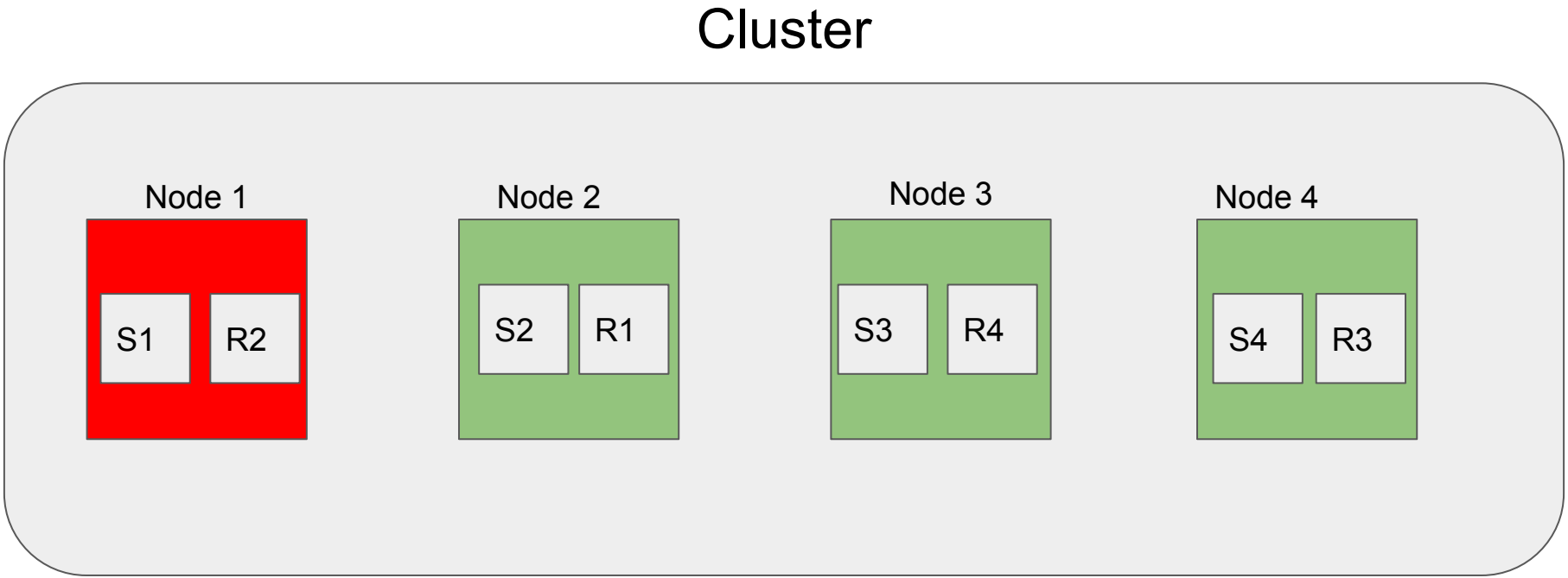
Cluster, Nodes, Shards and Replicas

Cluster, Nodes, Shards and Replicas

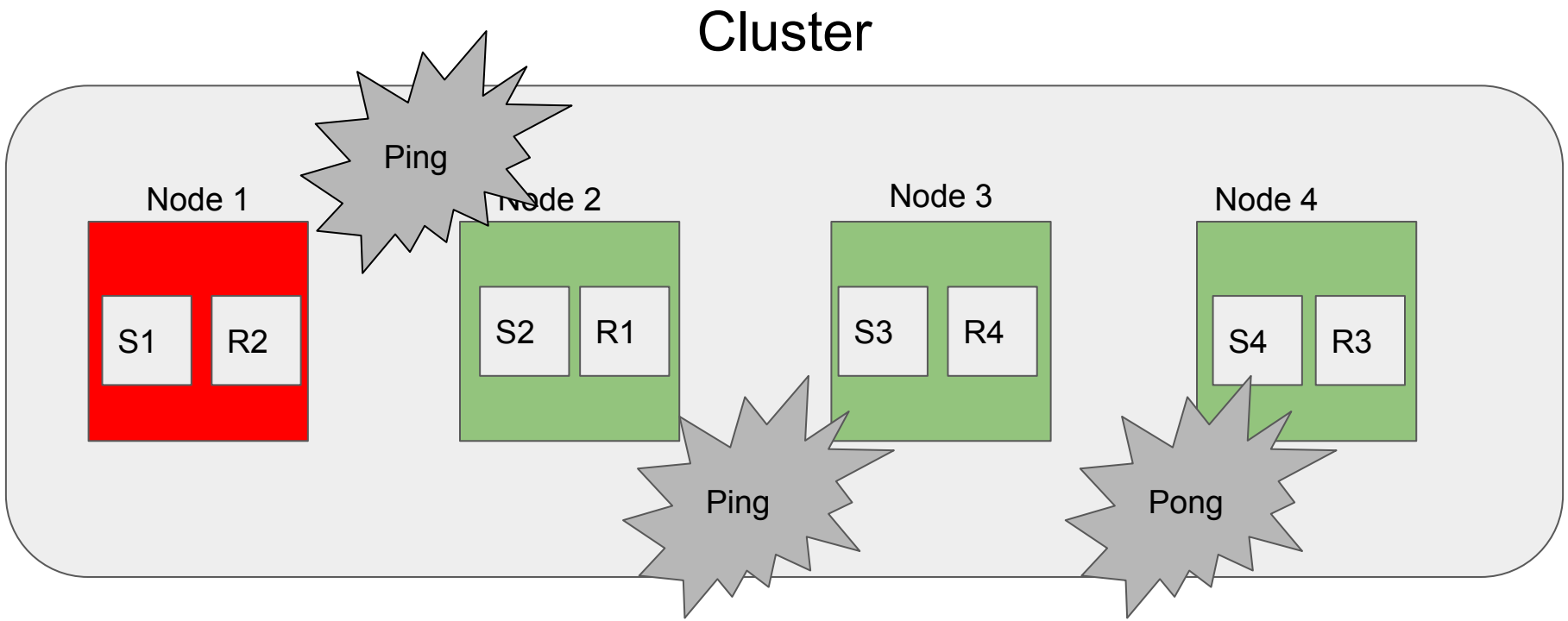
Cluster, Nodes, Shards and Replicas



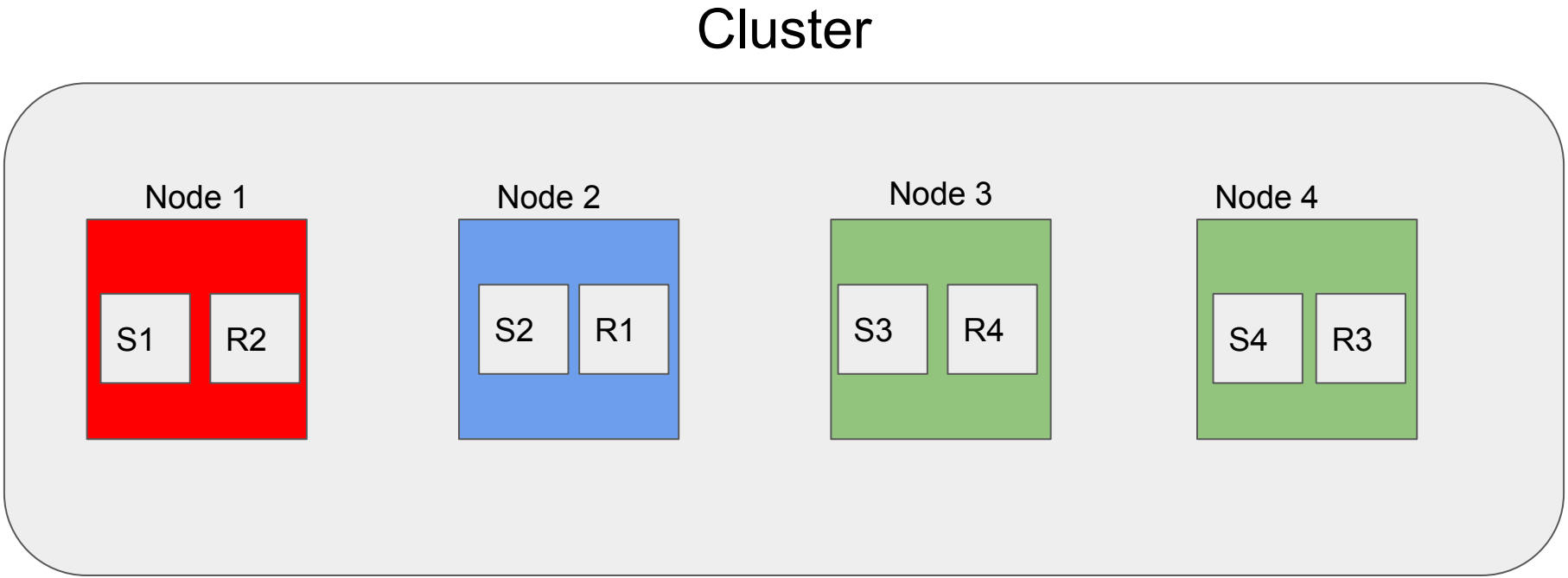
Cluster, Nodes, Shards and Replicas



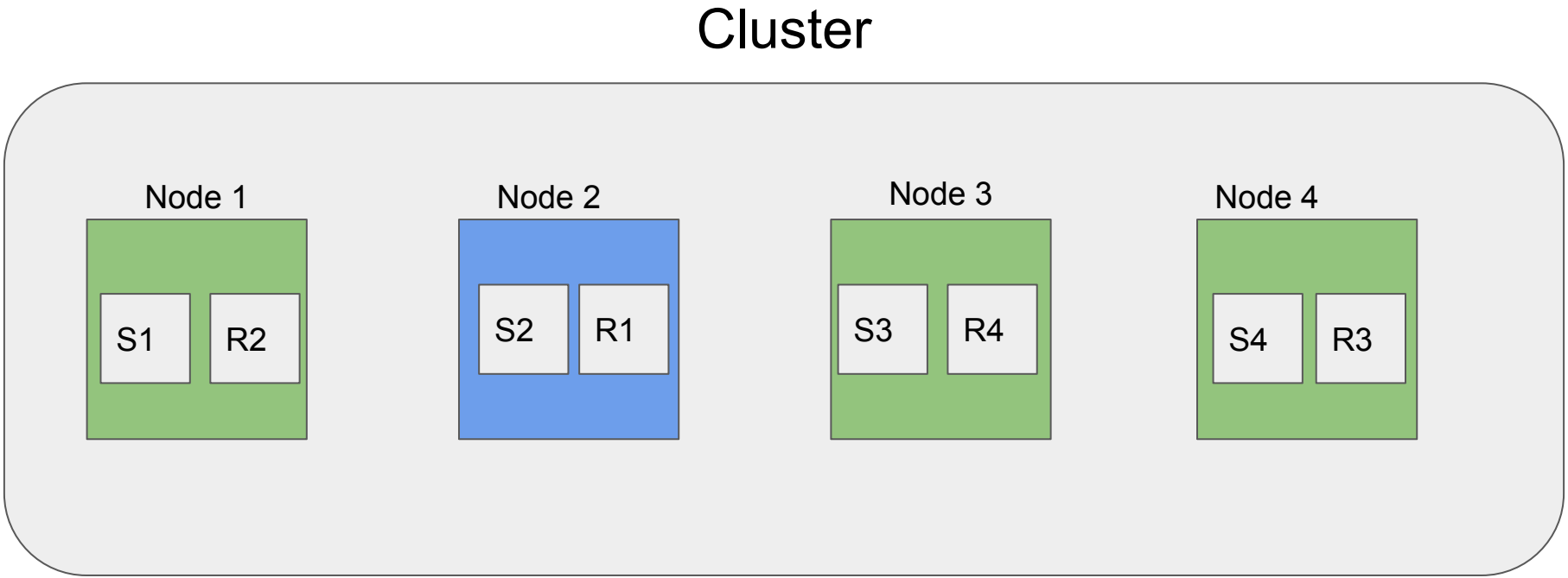
Cluster, Nodes, Shards and Replicas



Cluster, Nodes, Shards and Replicas



Cluster, Nodes, Shards and Replicas



Responsibilities of the master

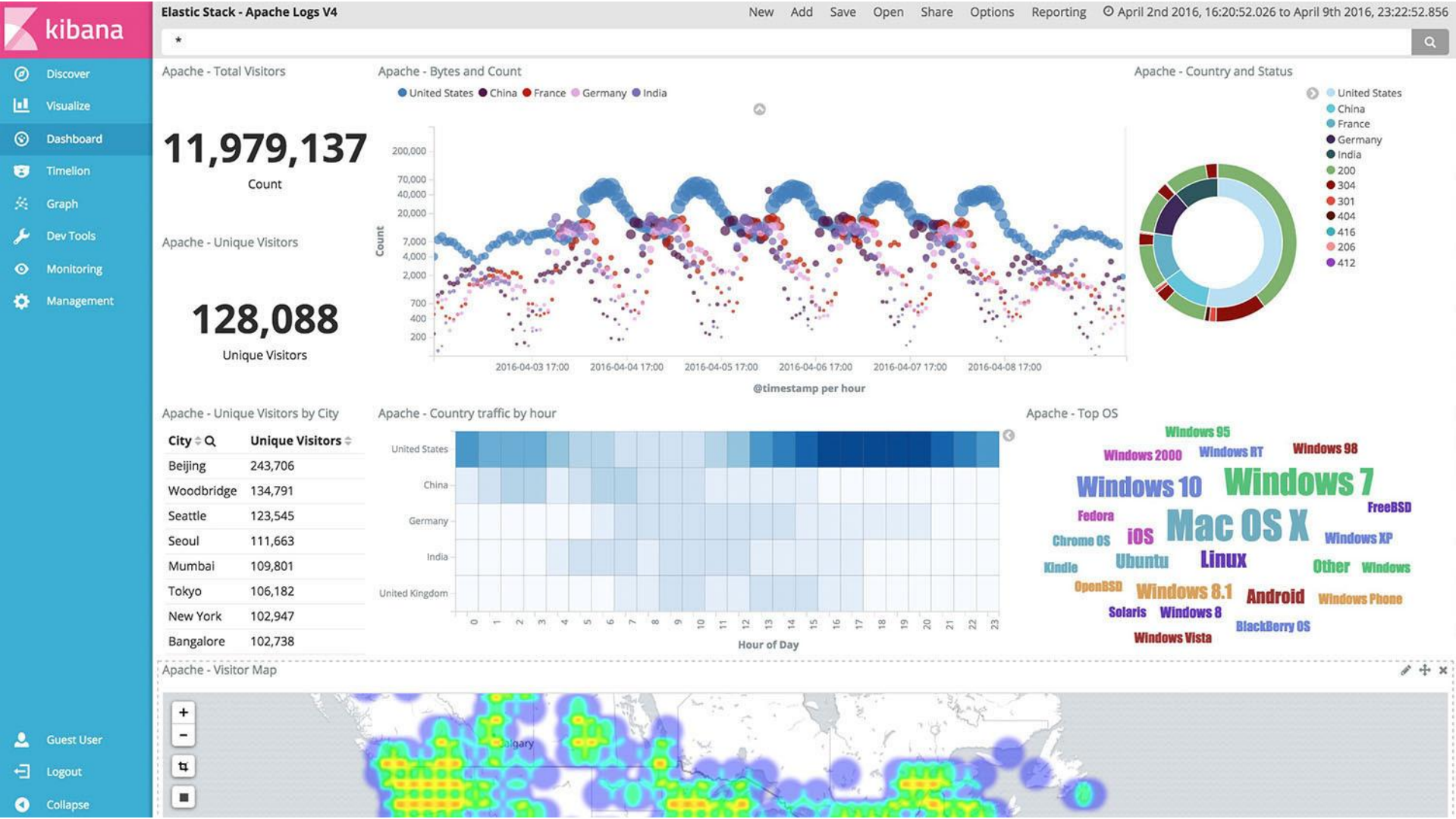
- Cluster health
- All the creation of index
- Repartition of the Shards
- Repartition of the Replicas

Cluster recommendation

- Your servers in the same data center
- Your machines on different Rack
- Keeping at least 3 eligible master node (Quorum of 2 is 2)

What’s Kibana?

- Another elastic product
- A tool allowing you to communicate in a more “human” way to your elasticsearch
- A product that allow you to do dashboard and data visualization


kibana
Elastic Stack - Apache Logs V4
New
Add
Save
Open
Share
Options
Reporting
April 2nd 2016, 16:20:52.026 to April 9th 2016, 23:22:52.856
*
Discover
Visualize
Dashboard
Timelion
Graph
Dev Tools
Monitoring
Management
Guest User
Logout
Collapse
Apache - Total Visitors
11,979,137
Count
Apache - Unique Visitors
128,088
Unique Visitors
Apache - Bytes and Count
United States
China
France
Germany
India
200,000
70,000
40,000
20,000
7,000
4,000
2,000
700
400
200
Count
2016-04-03 17:00
2016-04-04 17:00
2016-04-05 17:00
2016-04-06 17:00
2016-04-07 17:00
2016-04-08 17:00
@timestamp per hour
Apache - Country and Status
United States
China
France
Germany
India
200
304
301
404
416
206
412
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
United States
China
Germany
India
United Kingdom
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day
Apache - Top OS
Windows 95
Windows 2000
Windows RT
Windows 98
Windows 10
Windows 7
FreeBSD
Fedora
iOS
Mac OS X
Chrome OS
Windows XP
Kindle
Ubuntu
Linux
Other
Windows
OpenBSD
Windows 8.1
Android
Windows Phone
Solaris
Windows 8
BlackBerry OS
Windows Vista
Apache - Visitor Map
Calgary

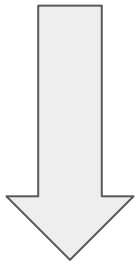
Let’s go for a demonstration

Demonstration done on Kibana

Query can be found on Github:

The analyzer

```
PUT product/book/0
{
  "title" : "A walk in the wood"
}
```

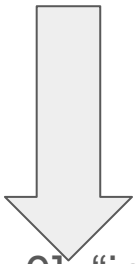


Standard Analyzer

{“a”: [id_0], “walk”: [id_0], “in”: [id_0], “the”: [id_0], “wood”: [id_0]}

The analyzer

```
PUT product/book/1
{
  "title" : "Probability: A complete guide"
}
```

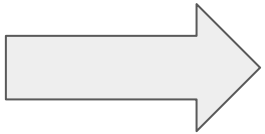


Standard Analyzer

{“a”: [id_0, id_1], “walk”: [id_0], “in”: [id_0], “the”: [id_0], “wood”: [id_0], “probability”:[id_1], “complete”:[id_1], “guide”:[id_1]}

The analyzer

```
GET product/book/_search
{
  "query": {
    "match": {
      "title": "A"
    }
  }
}
```

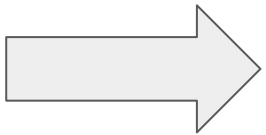


[id_0, id_1]

{“a”: [id_0, id_1], “walk”: [id_0], “in”: [id_0],
“the”: [id_0], “wood”: [id_0], “probability”:[id_1],
“complete”:[id_1], “guide”:[id_1]}

The analyzer

```
GET product/book/_search
{
  "query": {
    "term": {
      "title": {
        "value": "A"
      }
    }
  }
}
```

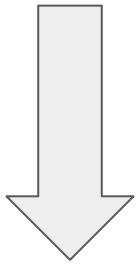


[]

{“a”: [id_0, id_1], “walk”: [id_0], “in”: [id_0],
“the”: [id_0], “wood”: [id_0],
“probability”:[id_1], “complete”:[id_1],
“guide”:[id_1]}

The english analyzer

```
PUT product/book/0
{
  "title" : "A walk in the wood"
}
```

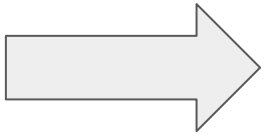


English Analyzer

{“walk”: [id_0], “wood”: [id_0]}

The english analyzer

```
GET product/book/_search
{
  "query": {
    "match": {
      "title": "A"
    }
  }
}
```



[]

{ “walk”: [id_0], “wood”: [id_0]}

What is relevance?

Two theories to know:

- Boolean model
- Space vector model

Boolean model

O0 = “Eric is ... always feeding”

O1 = “Jherez is ... with the friends”

….

O6 = “Manage Idea… to Melvyn)”

QT= {“lab”, “manager”} QO = “OR”

T = {t1:”lab”, t2:”manager”, t3:”Idea”, …, “t4”: feeding}

D = {D0, D1, …, D6}

D0 = {Eric, is, …, feeding}

D1 = {Jherez, is, …, friends}

D6 = {Manage, idea, …, Melvyn}

S1 = {D0, D1, D6}

S2 = {D0, D6}

SF = S1 $\cup$ S2 = S1

Space vector model

S1 = {D0, D1, D6}

T0 = D0 ∩ QT (“lab”, “manager”) ⇒ V0 = (L0, M0)

T1 = D1 ∩ QT (“lab”) ⇒ V1 = (L1, 0)

T6 = D6 ∩ QT (“lab”, “manager”) ⇒ V6 = (L6, M6)

Weight of a token in a document

- Term frequency

$$TF = \sqrt{Frequency}$$

- Inverse Document Frequency

$$IDF = 1 + \log(1/ (docFrequency + 1))$$

- Field length

$$FL = 1 / \sqrt{TokenInField}$$

$$Weight = TF \times IDF \times FL$$

Relevance

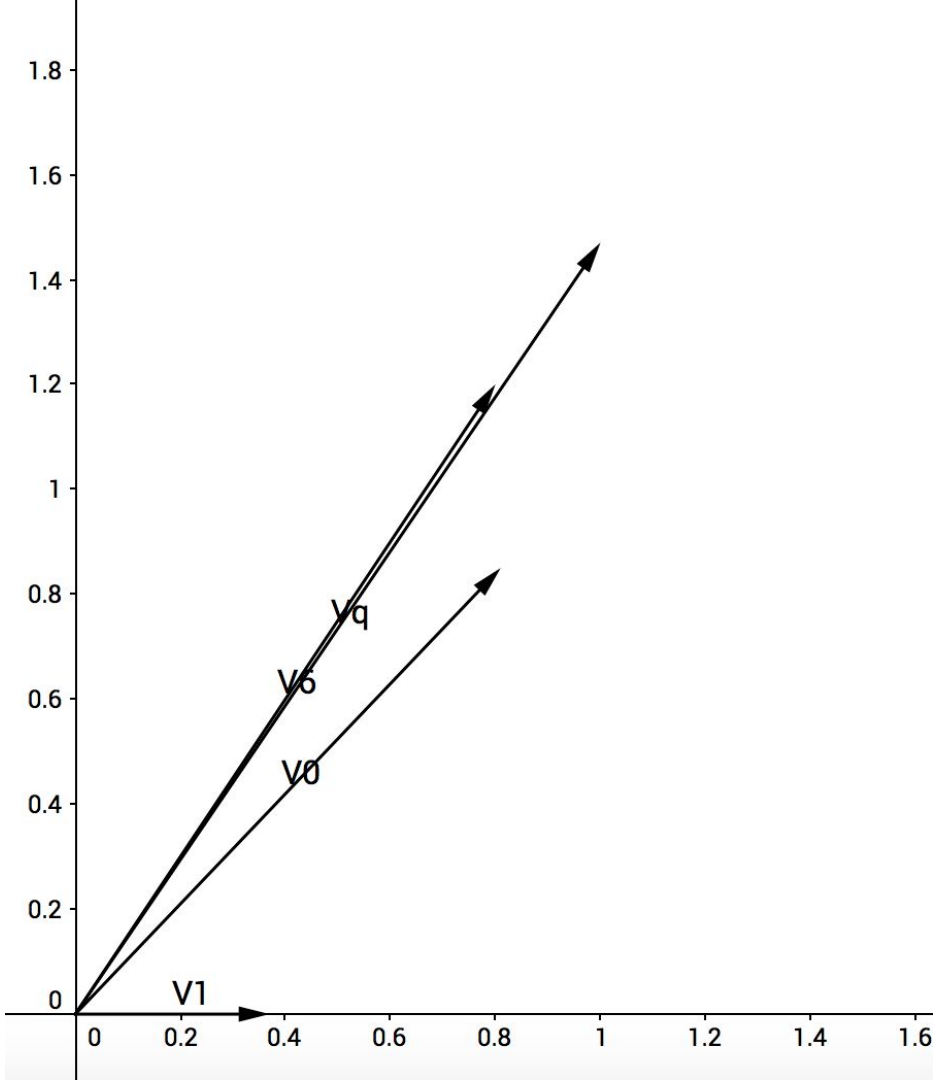
Vq = [1, 1.47]

V0 = [0.81, 0.85]

V1 = [0.37, 0]

V6 = [0.8, 1.2]

Relevance(Vq, Vx) = cos(Vq, Vx) = (Vq . Vx) / (‖Vq‖.‖Vx‖)



Let’s Kaggle with elasticsearch

https://www.kaggle.com/c/whats-cooking

Results of our “Classifier”

-	MelvynPeignon	0.76599	-	Wed, 22 Jun 2016 16:51:03	Post-Deadline

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

Explanation of the methodology:
http://melvyn.pythonanywhere.com/posts/1/

Last advices?

- Mapping (I highly recommend having a mapping. You cannot update the type defined in a field in the mapping)
- Elasticsearch as a database (I prefer having both, easier for reindexation, having a back up, do my search and analytics on ES and use my database for identification, etc ...)
- Elasticsearch as a NOSQL database (I wouldn’t do it on a serious project, but nice to have if you wanna do a quick implementation for a POC)

Hope you enjoyed the presentation!

Thank you for your attention!

Questions?