Elastic

Founded in 2012

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

What is elasticsearch?

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

Shay
Bannon

CRUD

CREATE

READ

UPDATE

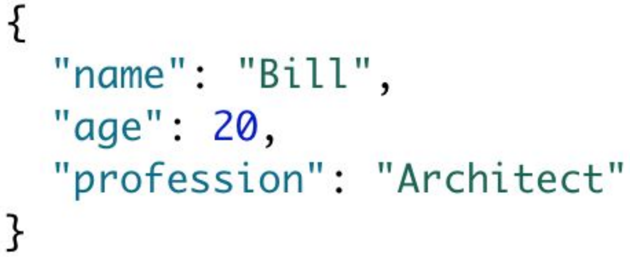
DELETE

Some concepts to know

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

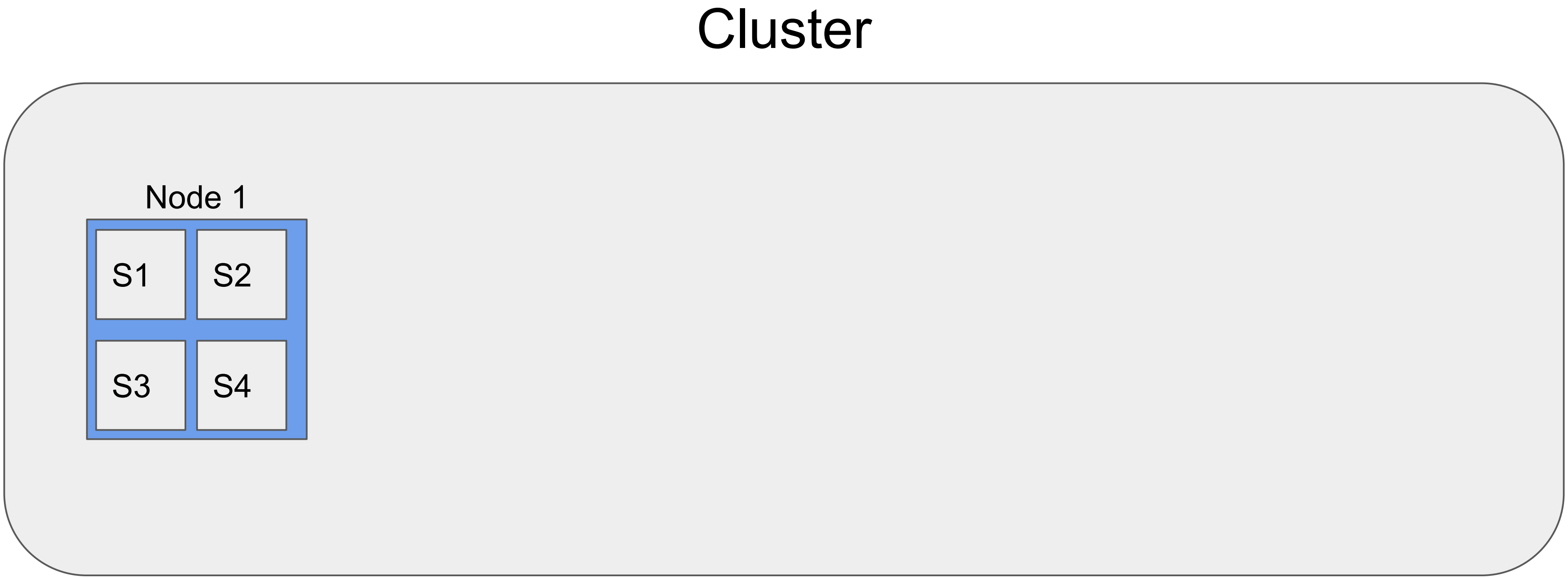
Documents, Types, indexes

- 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

Node 1

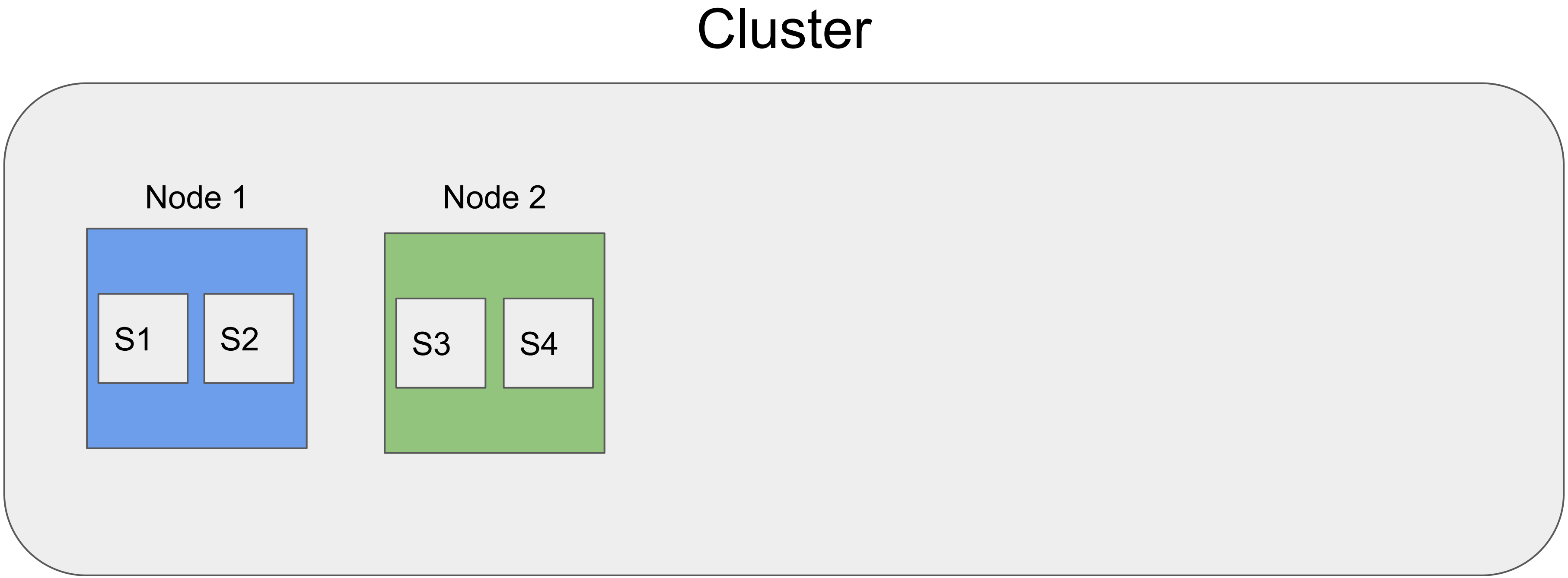
S1

S2

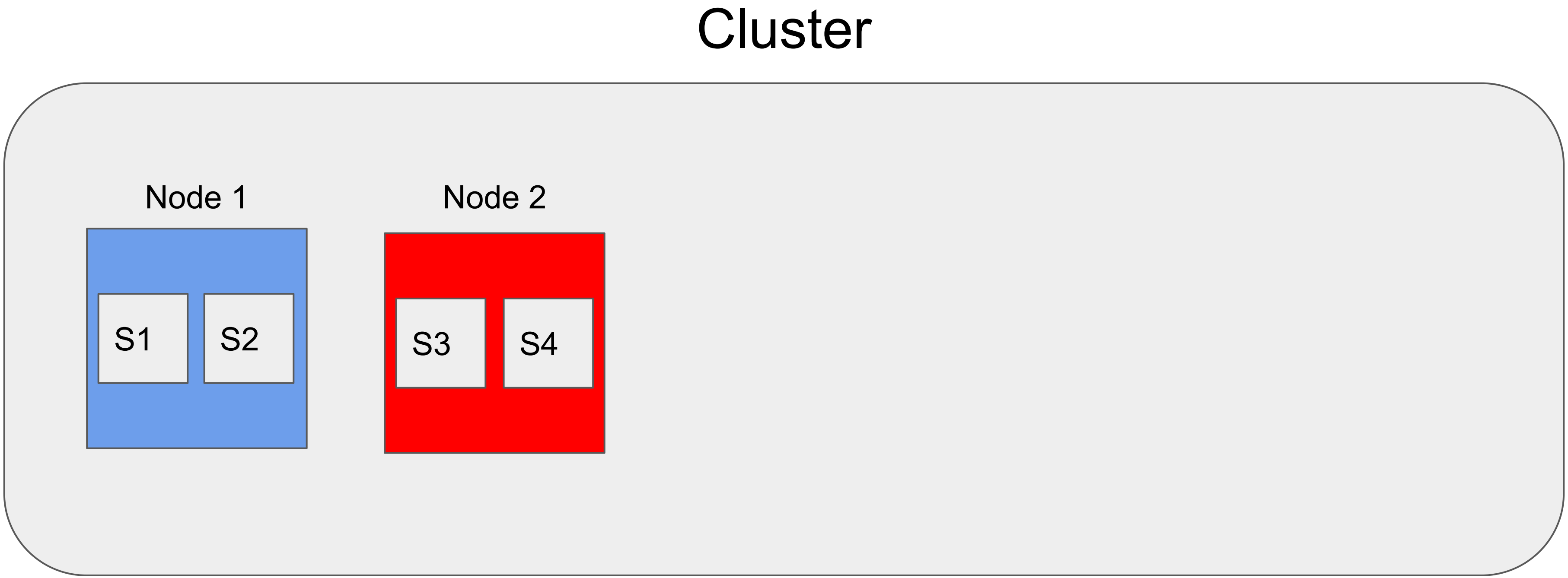
S3

S4

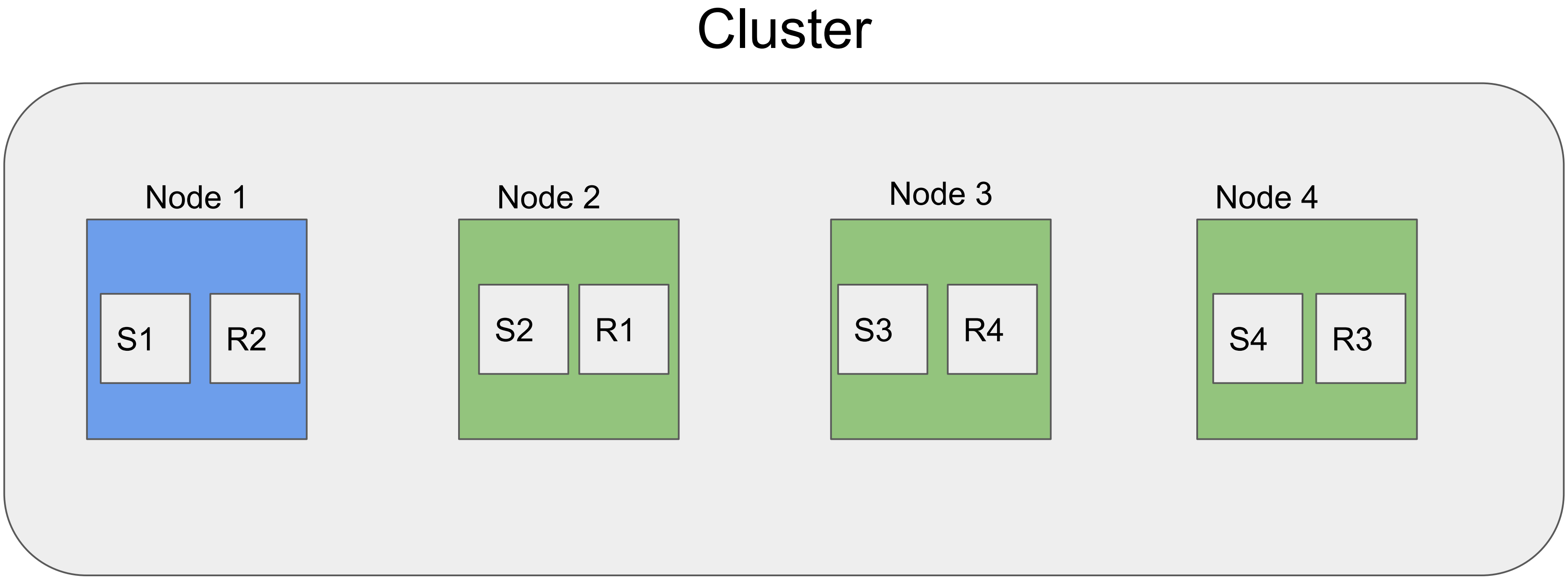
Cluster, Nodes, Shards and Replicas



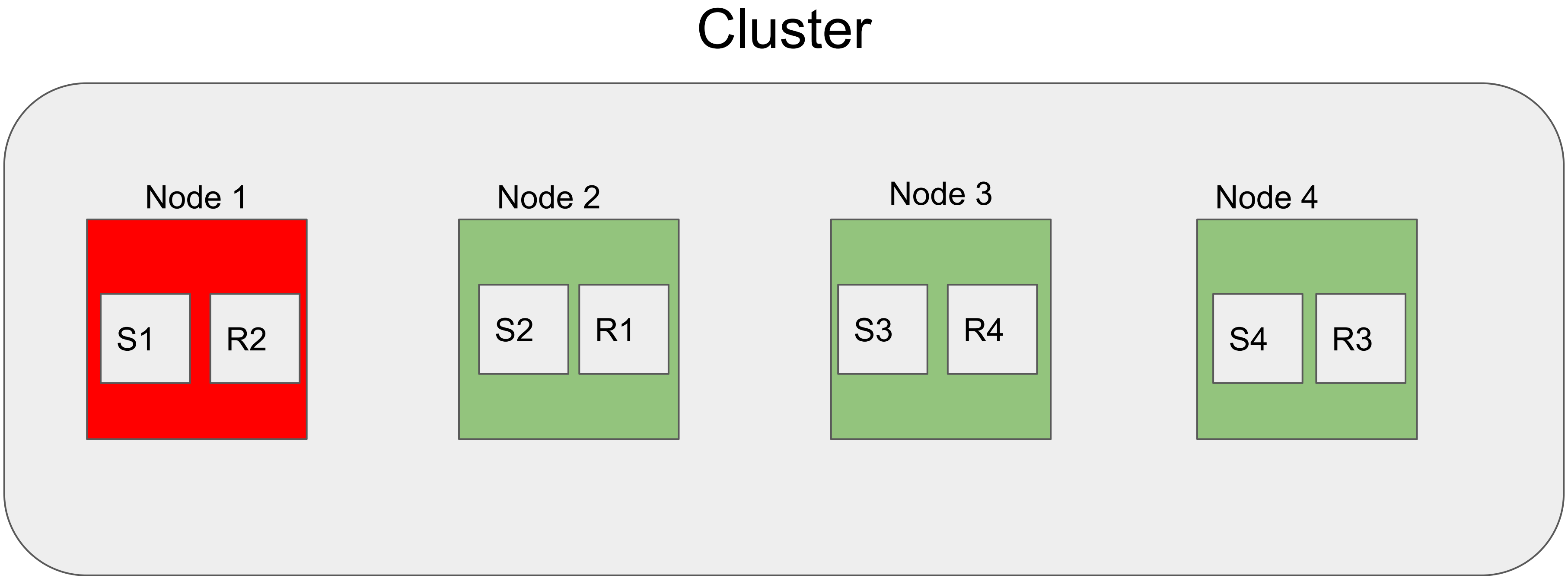
Cluster, Nodes, Shards and Replicas



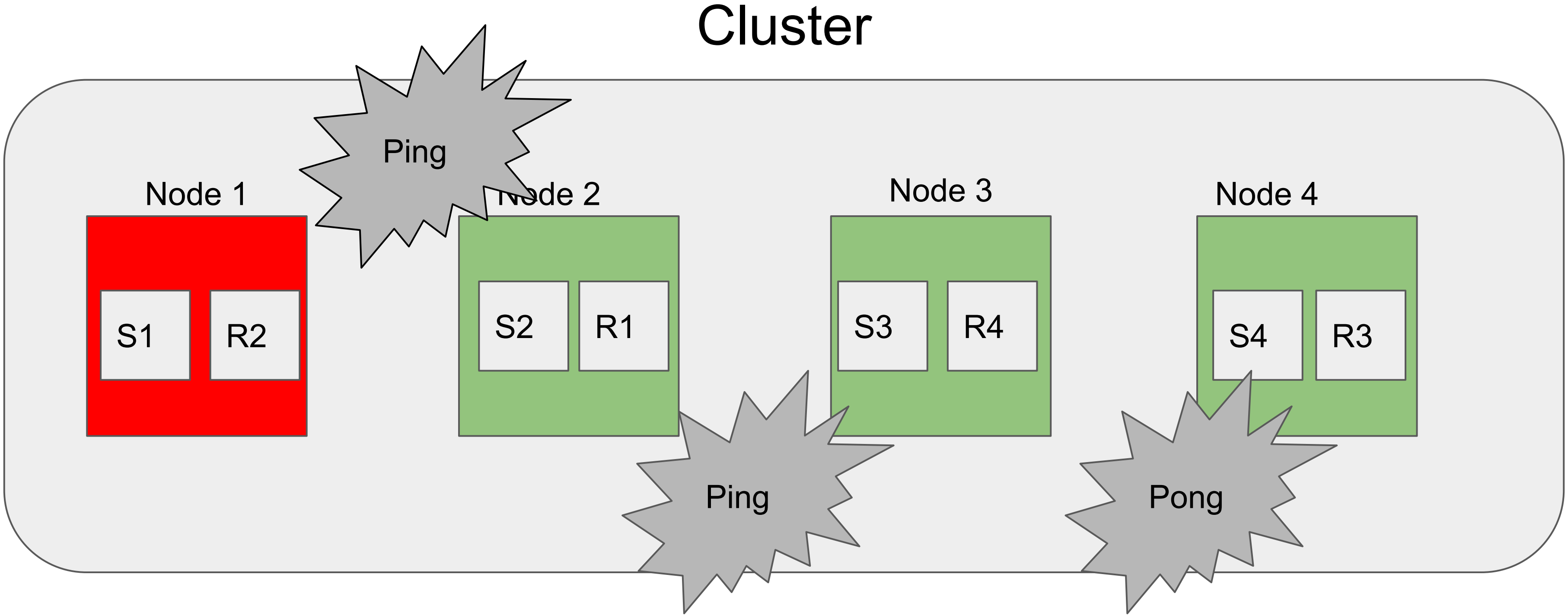
Cluster, Nodes, Shards and Replicas



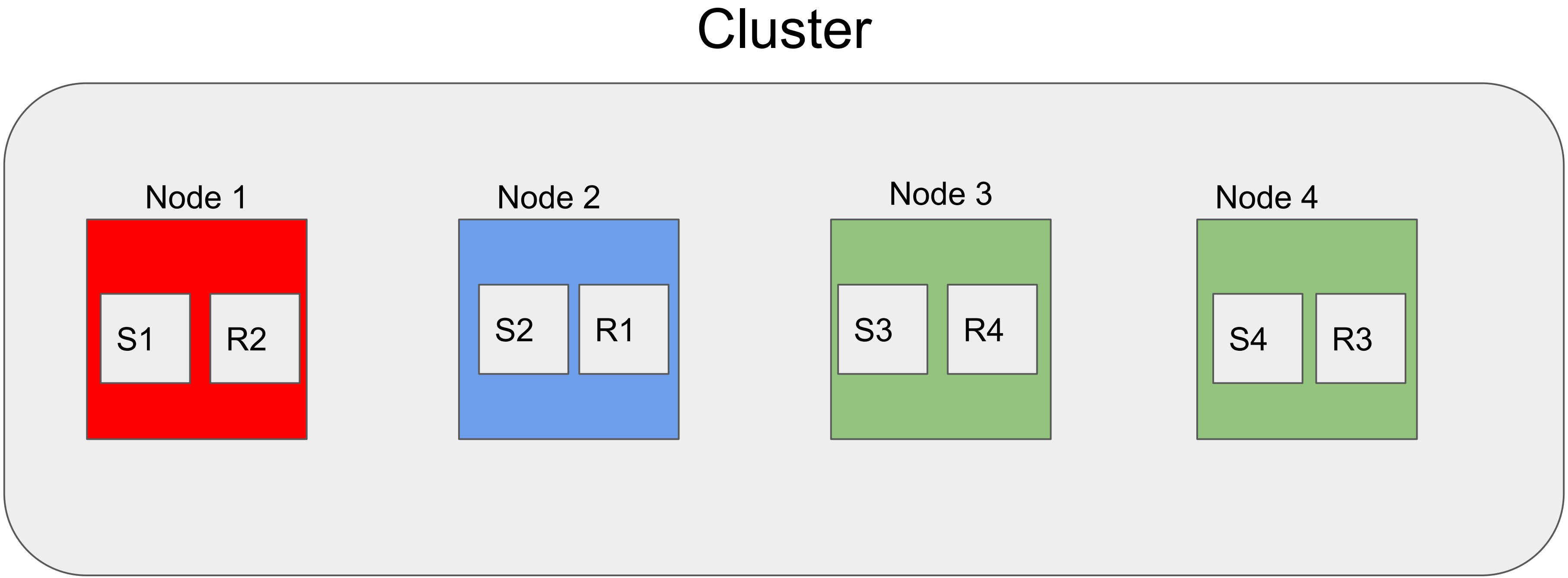
Cluster, Nodes, Shards and Replicas



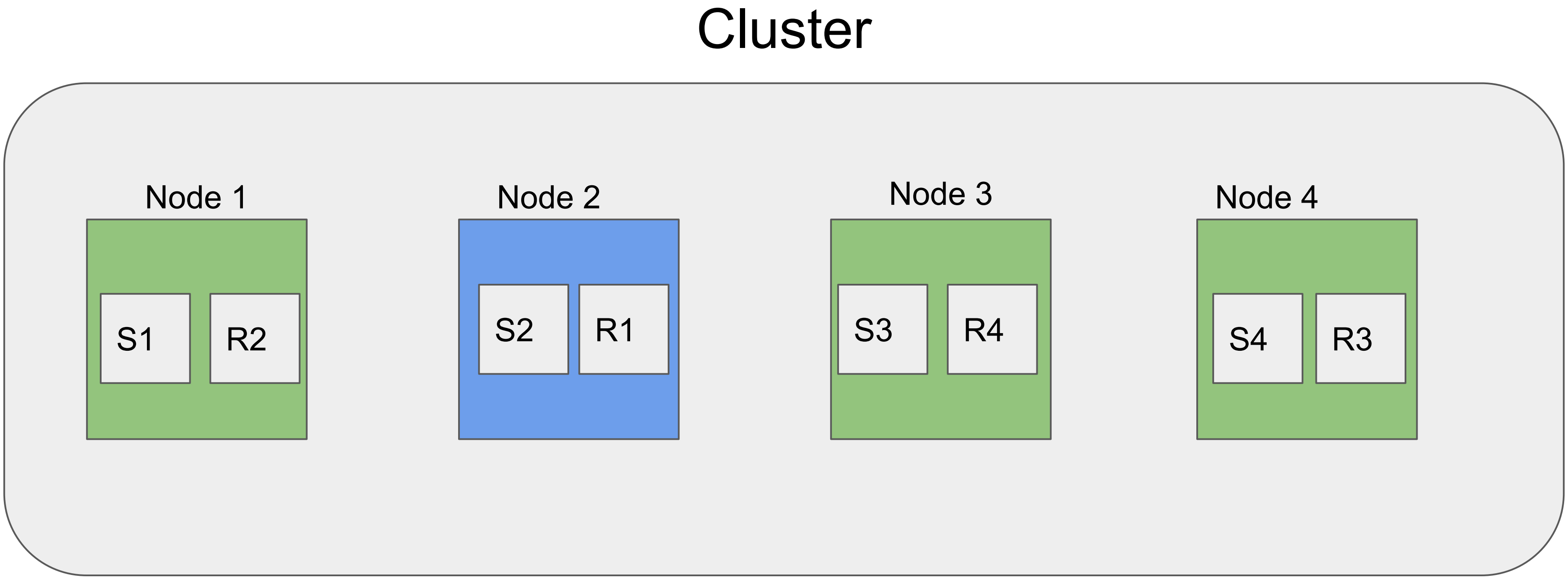
Cluster, Nodes, Shards and Replicas



Cluster, Nodes, Shards and Replicas



Cluster, Nodes, Shards and Replicas



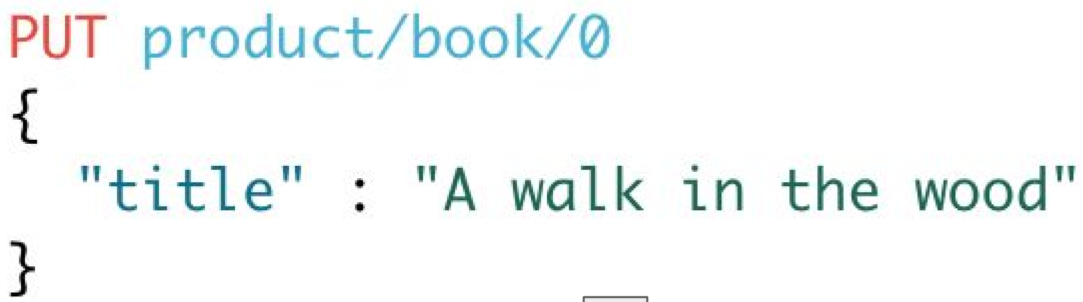
Responsibilities of the master

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

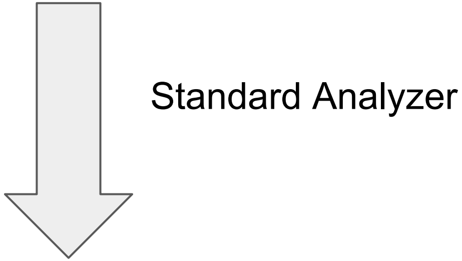
Cluster recommendation

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

The analyzer



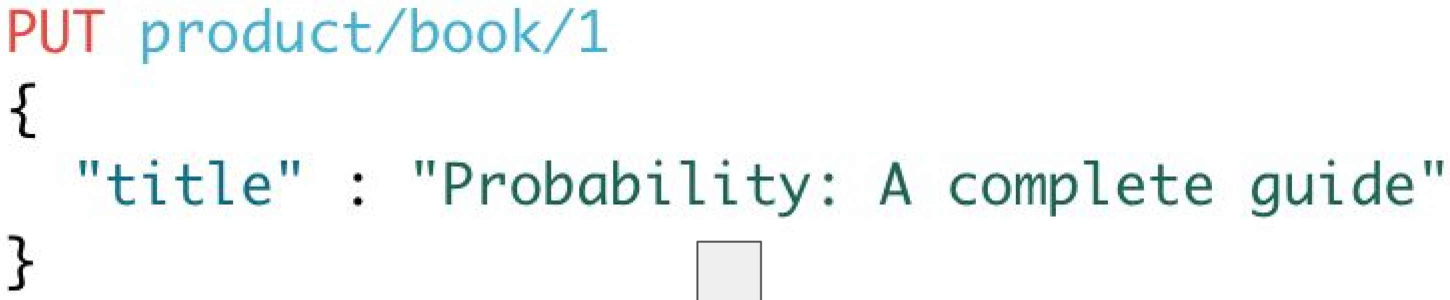
```
PUT product/book/0
{
  "title" : "A walk in the wood"
}
```



Standard Analyzer

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

The analyzer

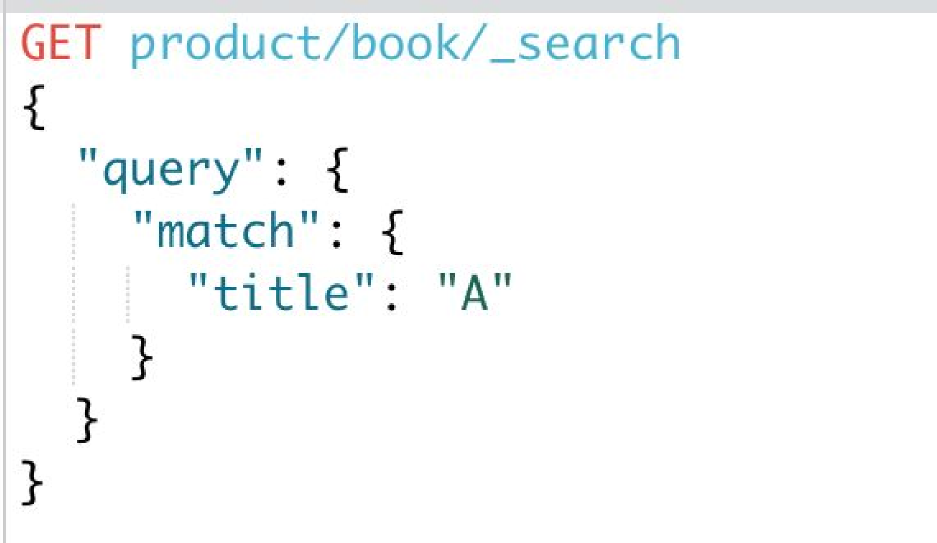


```
PUT product/book/1
{
  "title" : "Probability: A complete guide"
}
```

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

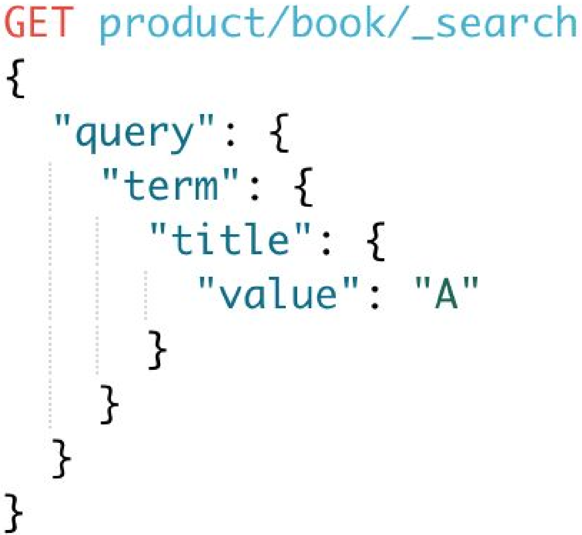


```
GET product/book/_search
{
  "query": {
    "match": {
      "title": "A"
    }
  }
}
```

[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

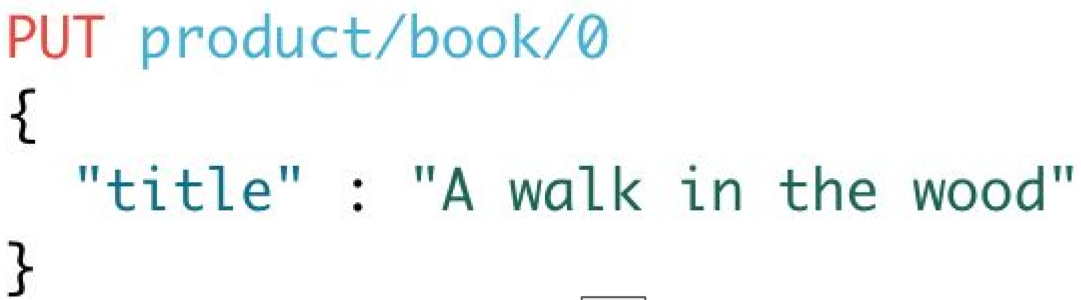


```
GET product/book/_search
{
  "query": {
    "term": {
      "title": {
        "value": "A"
      }
    }
  }
}
```

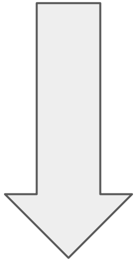
[]

{“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



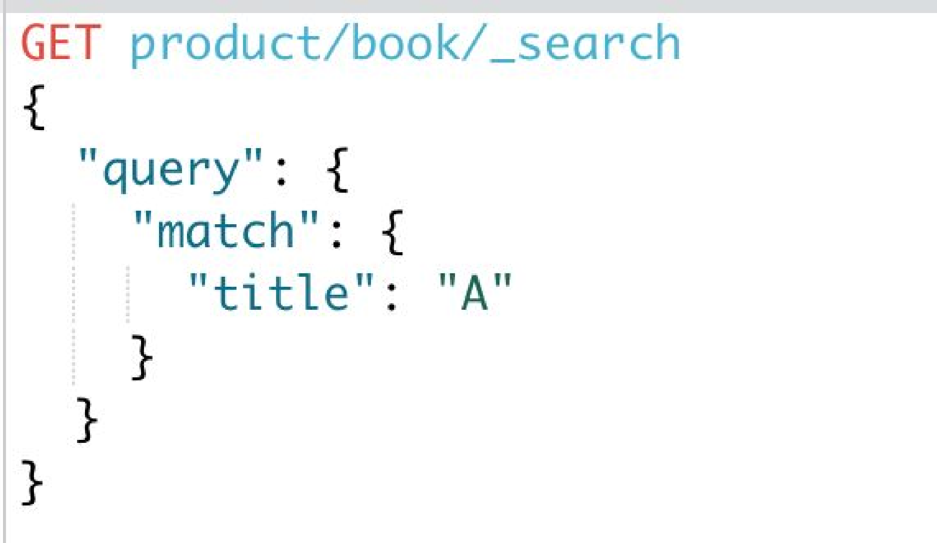
```
PUT product/book/0
{
  "title" : "A walk in the wood"
}
```



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 ∪ 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

$$\text{TF} = \sqrt{\text{Frequency}}$$

- Inverse Document Frequency

$$\text{IDF} = 1 + \log(1 / (\text{docFrequency} + 1))$$

- Field length

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

$$\text{Weight} = \text{TF x IDF x FL}$$

Relevance

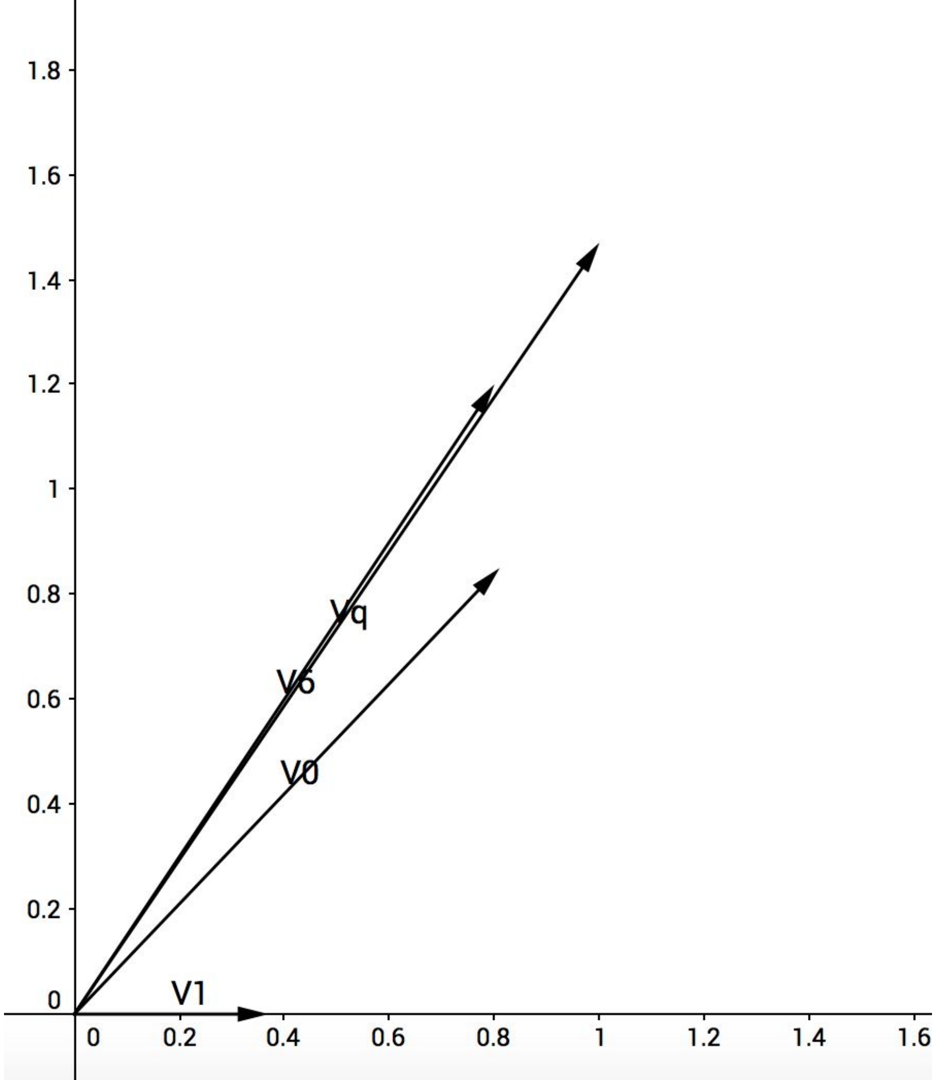
Vq = [1, 1.47]

V0 = [0.81, 0.85]

V1 = [0.37, 0]

V6 = [0.8, 1.2]

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


Vq
V6
V0
V1