

26/02/2015

Elasticsearch : Quick overview

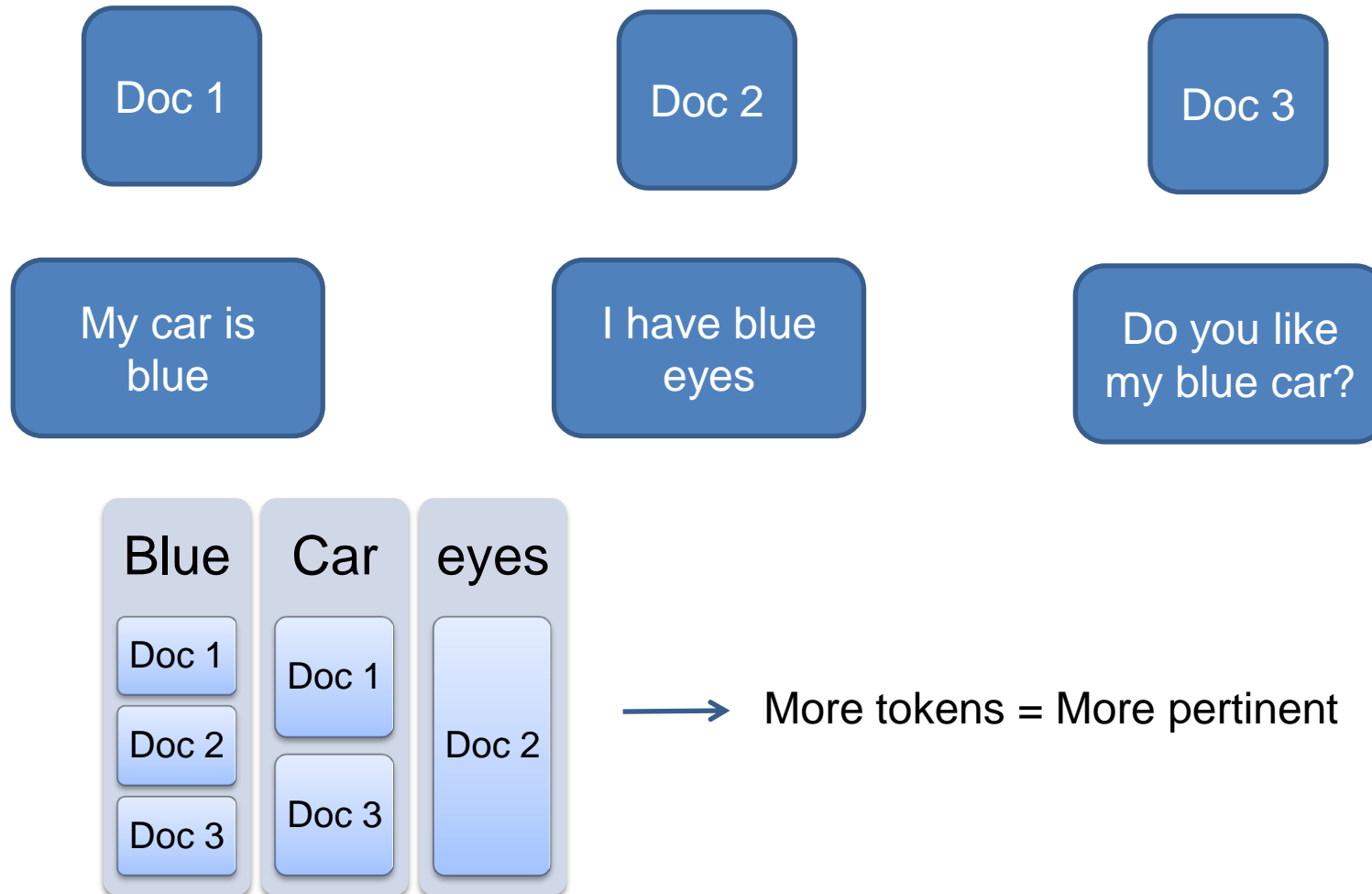
Business & Decision

| Introduction

Search Index

- Increase search performances
- Without index, we need to search documents in data-store
- Store document's words & place it in an inversed index

Search Index



Elasticsearch

- Based on Apache Lucene
- Scalable on hundred servers
- Real time Search engine
- API RESTFul

| Installation

Installation

Install Java 7 or Java 8

Download Zip : <http://www.elasticsearch.org/overview/elkdownloads/>

Run "es_folder/bin/elasticsearch -d"

Open "localhost:9200" in a browser and you should see

```
{
  "status" : 200,
  "name" : "Blackwing",
  "version" : {
    "number" : "1.3.4",
    "build_hash" : "a70f3ccb52200f8f2c87e9c370c6597448eb3e45",
    "build_timestamp" : "2014-09-30T09:07:17Z",
    "build_snapshot" : false,
    "lucene_version" : "4.9"
  },
  "tagline" : "You Know, for Search"
}
```

Plugins

- Monitoring
- Indexation, OCR, image...
- ICU (non-english languages)
- Inspection, development
- Aggregations, scripting, etc.

Plugins: Head

The screenshot shows the ElasticSearch Head interface for a cluster named 'Savage Steel'. The cluster health is 'red (1, 63)'. The page displays a table of indices and their shard distribution across nodes.

| Index | Size | Docs | Shard 0 | Shard 1 | Shard 2 | Shard 3 | Shard 4 |
|-------------|-------------|-------|---------|---------|---------|---------|---------|
| beer | 495b | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| beer-sample | 198b (198b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| dfgd | 330b (330b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| er | 495b (495b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| gagan | 495b (495b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| gh | 495b (495b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| google | 330b (330b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| mahesh | 330b (330b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| sandeep | 330b (330b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| user | 495b (495b) | 0 (0) | 0 | 1 | 2 | 3 | 4 |
| Unassigned | | | 0 | 1 | 2 | 3 | 4 |

Plugins: Kopf

The screenshot displays the Kopf cluster administration interface. At the top, a green header bar contains the 'kopf' logo, navigation links for 'cluster', 'rest', and 'plugins', and a 'connect to a different host' button. Below the header, the 'CLUSTER ADMINISTRATION' section shows cluster statistics: 2 nodes, 4 indices, 40 shards, 0 failed shards, 0 unassigned shards, 0 docs, and 3.48KB total size. A filter bar allows selecting by index type (data, master, shard) and shows '1-4 of 4' items. The main table lists indices: 'Hypokel' and 'Anything', each with 5 shards. The table is organized into columns for different indices: actions, companies, logging, and users. Each cell in the table shows a grid of green squares representing shards.

| | actions shards: 5 + 1 docs: 0 size: 495.00B | companies shards: 5 + 1 docs: 0 size: 495.00B | logging shards: 5 + 1 docs: 0 size: 495.00B | users shards: 5 + 1 docs: 0 size: 495.00B |
|---|--|--|--|--|
| Hypokel map(100, 100, 1, 01, 0002) index(100, 100, 1, 01, 0002) size: 0.00 heap: 01.000000000B | | | | |
| Anything map(100, 100, 1, 01, 0002) index(100, 100, 1, 01, 0002) size: 0.00 heap: 01.000000000B | | | | |
| unassigned shards | | | | |

Plugins : ElasticHQ

The screenshot displays the ElasticHQ web interface. At the top, there's a header with the ElasticHQ logo, a connection URL (http://192.168.73.128:9200), and a 'Connect' button. To the right are links for 'My Settings', 'Get Help', and 'Star us on GitHub'. Below the header, there's a navigation bar with 'elasticsearch' and 'Node Diagnostics' buttons on the left, and 'Indices', 'Query', 'Mappings', and 'REST' tabs on the right. The main content area is titled 'Cluster Overview' with a timestamp of 09:11:05. Below this is a 'Cluster Statistics' section with six cards showing: 1 Node, 20 Total Shards, 10 Successful Shards, 2 Indices, 3 Total Documents, and 8kb Total Size. At the bottom, there are two sections: 'Cluster Health' and 'Indices'. The 'Cluster Health' section shows a 'Yellow' status and a table with details like 'Timed Out?' (false), '# Nodes' (1), '# Data Nodes' (1), and 'Active Primary Shards' (10). The 'Indices' section shows a table with two indices: 'plants' and 'fibers', each with 2 and 1 documents respectively, 3.9kb and 4kb primary sizes, 5 shards each, 1 replica each, and 'open' status.

Elastic HQ <http://192.168.73.128:9200> [Connect](#) [My Settings](#) [Get Help](#) [Star us on GitHub](#)

[elasticsearch](#) [Node Diagnostics](#) [Baron Macabre](#) [Indices](#) [Query](#) [Mappings](#) [REST](#)

09:11:05 Cluster Overview

Cluster Statistics

| | | | | | |
|-------------------|---------------------------|--------------------------------|---------------------|-----------------------------|--------------------------|
| 1 Nodes | 20 Total Shards | 10 Successful Shards | 2 Indices | 3 Total Documents | 8kb Total Size |
|-------------------|---------------------------|--------------------------------|---------------------|-----------------------------|--------------------------|

Cluster Health

| | |
|-----------------------|--------|
| Status | Yellow |
| Timed Out? | false |
| # Nodes | 1 |
| # Data Nodes | 1 |
| Active Primary Shards | 10 |

Indices

| Index | # Docs | Primary Size | # Shards | # Replicas | Status |
|------------------------|--------|--------------|----------|------------|--------|
| plants | 2 | 3.9kb | 5 | 1 | open |
| fibers | 1 | 4kb | 5 | 1 | open |

Plugins : Marvel (official)



Plugins : Management

Install a plugin : `/bin/plugin --install mobz/elasticsearch-head`

Remove a plugin : `./bin/plugin --remove mobz/elasticsearch-head`

List plugins : `./bin/plugin --list`

Warning : No update possible (remove/ install only)

Official list available here :

<http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/modules-plugins.html>

Business & Decision

| Life in cluster

Life in cluster: Key points

A cluster is a group of nodes

A node is an Elasticsearch instance (One per machine)

There is automatically one master node

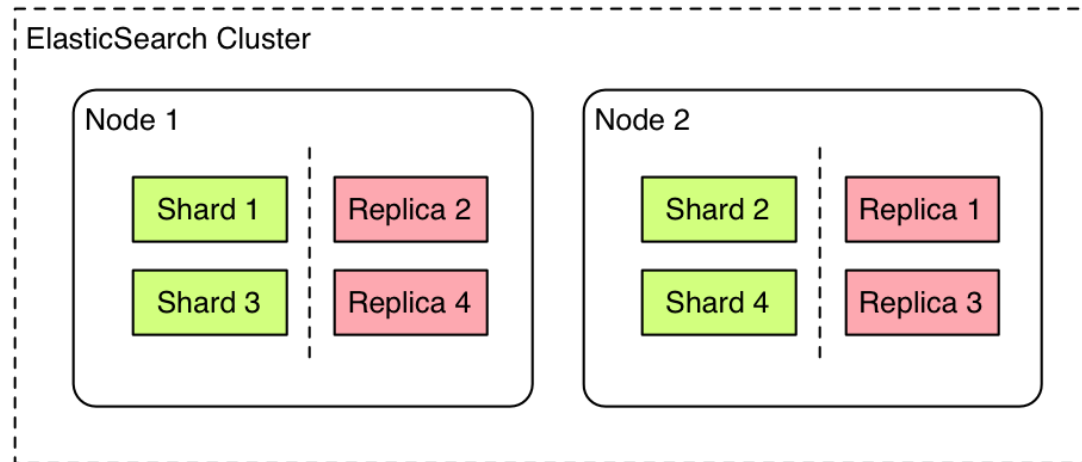
A node is composed by a fixed number of shards

Life in cluster : Key points

A replica is a copy of a shard

Replicas can be increased without re-indexing

A shard is a Lucene instance & contains documents



| Search / Index API

Create / Delete an index

Create an index : POST /my_index

Get info from an index : GET /my_index

Remove an index : DELETE /my_index

Create a document

Simple query

```
GET /wikipedia/_search
{
  "query": {
    "match": {
      "text": "bacterial"
    }
  }
}
```

Same with query string : GET /wikipedia/_search?q=text:bacterial

```
GET /wikipedia/_search
{
  "query": {
    "match_phrase": {
      "text": "bacterial chromosome"
    }
  }
}
```

Filtered query

Search for "bacterial" & filter by stub=false

```
GET /wikipedia/_search
{
  "query": {
    "filtered": {
      "filter": {
        "term": {
          "stub": false
        }
      },
      "query": {
        "match": {
          "title": "bacterial"
        }
      }
    }
  }
}
```

Filter by ID

```
GET /wikipedia/_search
{
  "query": {
    "filtered": {
      "filter": {
        "term": {
          "_id": "1628"
        }
      }
    }
  }
}
```

bool query

Search for documents (10 < docs < 20) about history & astronomy with a boost for documents about meteor Sorting by ID desc

```
GET /wikipedia/_search
{
  "query": {
    "bool": {
      "must": {"match": {"text": "history"}},
      "should": {"match": {"title": "astronomy"}},
      "should": {"match": {"text": "meteor^2"}}
    }
  },
  "sort": [
    {
      "_id": {
        "order": "desc",
      },
    }
  ],
  "size": 10,
  "from": 10
}
```

| Mapping

Mapping

Create a schema for you index

Indicate data types, analyzers , index action, etc.

Mapping cannot be updated without re-indexing

Mapping

```
POST /wiki
{
  "mappings": {
    "page": {
      "_all": {
        "enabled": false
      },
      "_size": {
        "enabled": true,
        "store": true
      },
      "properties": {
        "category": {
          "type": "string",
          "index": "not_analyzed"
        },
        "link": {
          "type": "string"
        },
        "stub": {
          "type": "boolean"
        },
        "text": {
          "type": "string",
          "analyzer": "french"
        },
        "title": {
          "type": "string",
          "analyzer": "french"
        }
      }
    }
  },
  "settings": {
    "number_of_shards": 2,
    "number_of_replicas": 1
  }
}
```

| Analyzers

Analyzer

Create tokens from words to increase search

3 steps :

- characters filter (html strip, clean text, stop words, etc.)
- Tokenizer (split text into tokens)
- Token filter : Transform tokens (lowercase, etc.)

Many defaults analyzers (whitespace, standard, simple, etc.) :

<http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/analysis-analyzers.html>

Analizers

Define Analyzers, filters, tokenizers

```
PUT /my_index
{
  "settings": {
    "analysis": {
      "char_filter": { ... custom character filters ... },
      "tokenizer": { ... custom tokenizers ... },
      "filter": { ... custom token filters ... },
      "analyzer": {
        "my_analyzer": {
          "type": "custom",
          "char_filter": [ "html_strip", "my_char" ],
          "tokenizer": "standard",
          "filter": [ "lowercase", "my_stopwords" ]
        }
      }
    }
  }
}
```

And use them :

```
PUT /wiki
{
  "mappings": {
    "page": {
      "properties": {
        "title": {
          "type": "string",
          "analyzer": "my_analyzer"
        }
      }
    }
  }
}
```

| Aggregations

Aggregations

Get category terms count (10 first) & links count by category (10 first)

```
GET /wikipedia/_search
{
  "aggs": {
    "category_count": {
      "terms": {
        "field": "category",
        "size": 10
      },
    },
    "aggs": {
      "link_count": {
        "terms": {
          "field": "link",
          "size": 10
        }
      }
    }
  }
}
```

| Relations

Relationnal

Parent / Child relations

- A relation between documents
- Updating documents separately

Nested

- Embed relations into a document (Ex: comments for a post)
- Can be used for query & filter

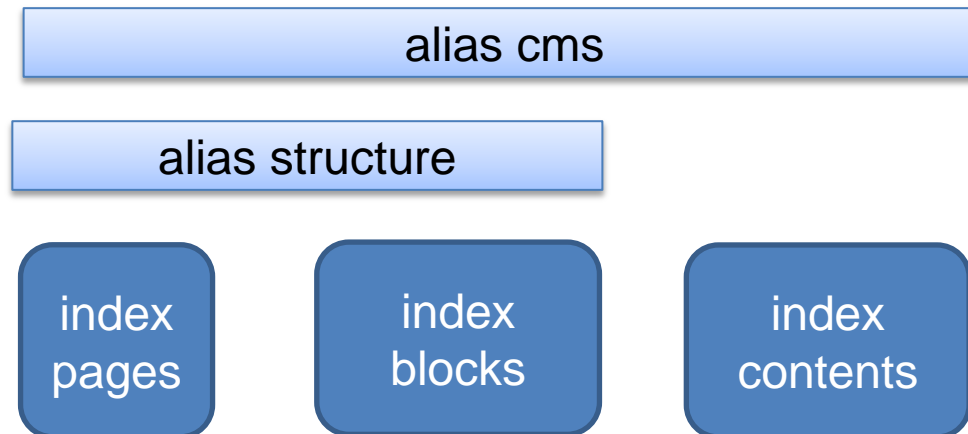
| Production

Alias

Don't use index directly

Avoid downtime when re-indexing

An alias can be a specific view (group of index)



Alias

Create index foo : "POST /foo"

Create alias foo_alias for index foo : "PUT /foo/_alias/foo_alias"

Create index new_foo : "POST /new_foo"

Switch alias :

```
POST /_aliases
{
  "actions": [
    { "remove":
      { "index": "foo", "alias": "foo_alias" }
    },
    { "add":
      { "index": "new_foo", "alias": "foo_alias" }
    }
  ]
}
```

Backup / Monitoring

Monitoring :

```
* GET /_status
* GET /_cluster/health
```

Backup : * PUT /_snapshot/my_backup

```
{
  "type": "fs", # AWS / Azure are availables
  "settings": {
    "location": "/mount/backups/my_backup",
    "compress": true
  }
}
```

Business & Decision

| Demo

| Documentation

Documentation

Reference : <http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/index.html>

The book : <http://www.elasticsearch.org/guide/en/elasticsearch/guide/current/>

Presentation : <https://github.com/pdenis/elasticsearch-quicktour>

Business & Decision

| Questions ?