Business & Decision

26/02/2015

Elasticsearch : Quick overview

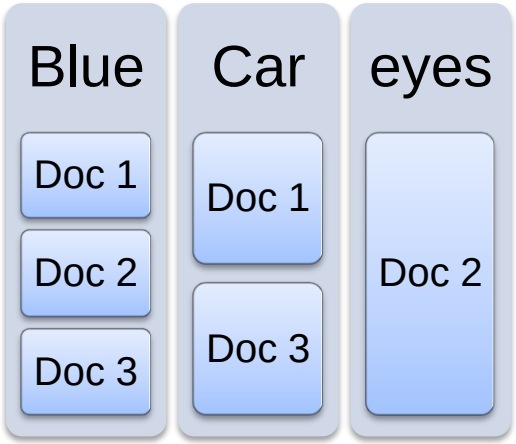
Business & Decision

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



→ More tokens = More pertinent

Elasticsearch

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

Business & Decision

Installation

Installation

Install Java 7 or Java 8

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

Run "es_folder/bin/elasticsearch -d"

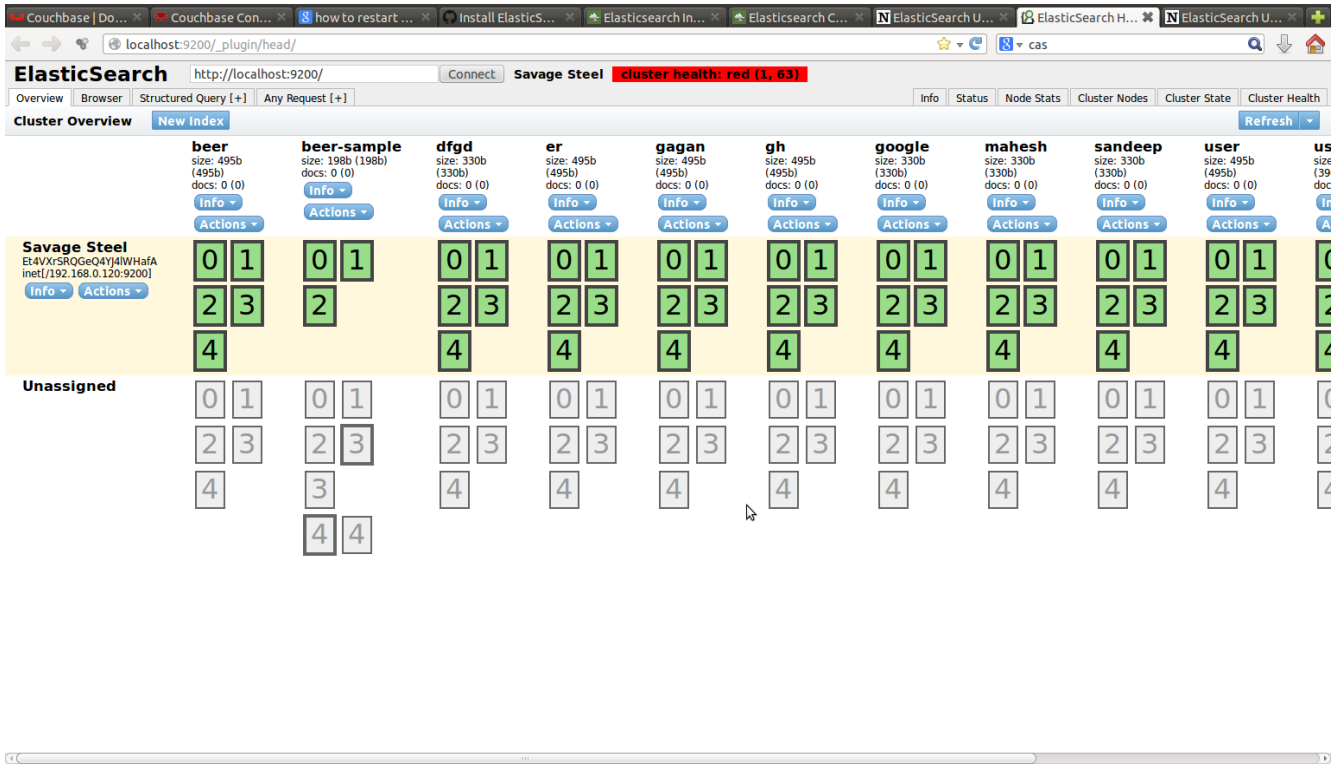
Open "localhost:9200" in a browser and you should see


```
{
  "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"
}
```

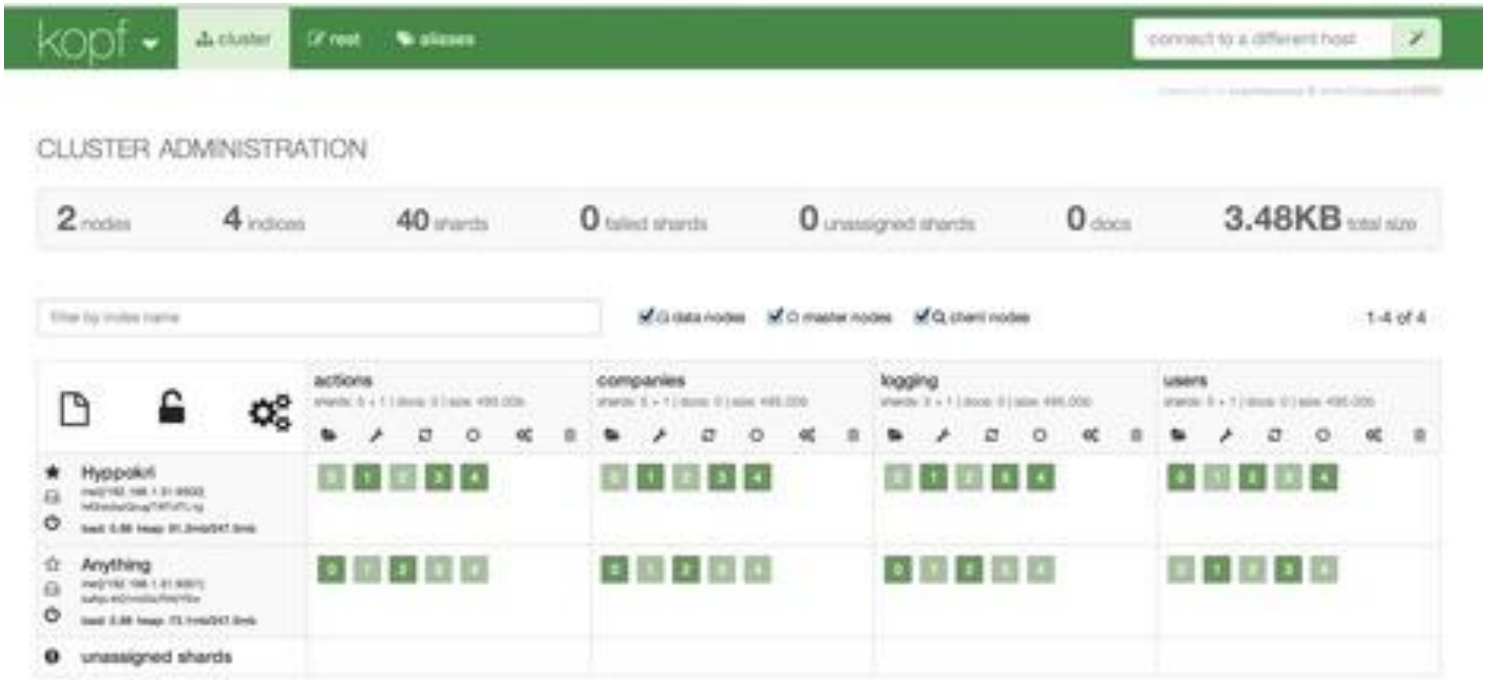
Plugins

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

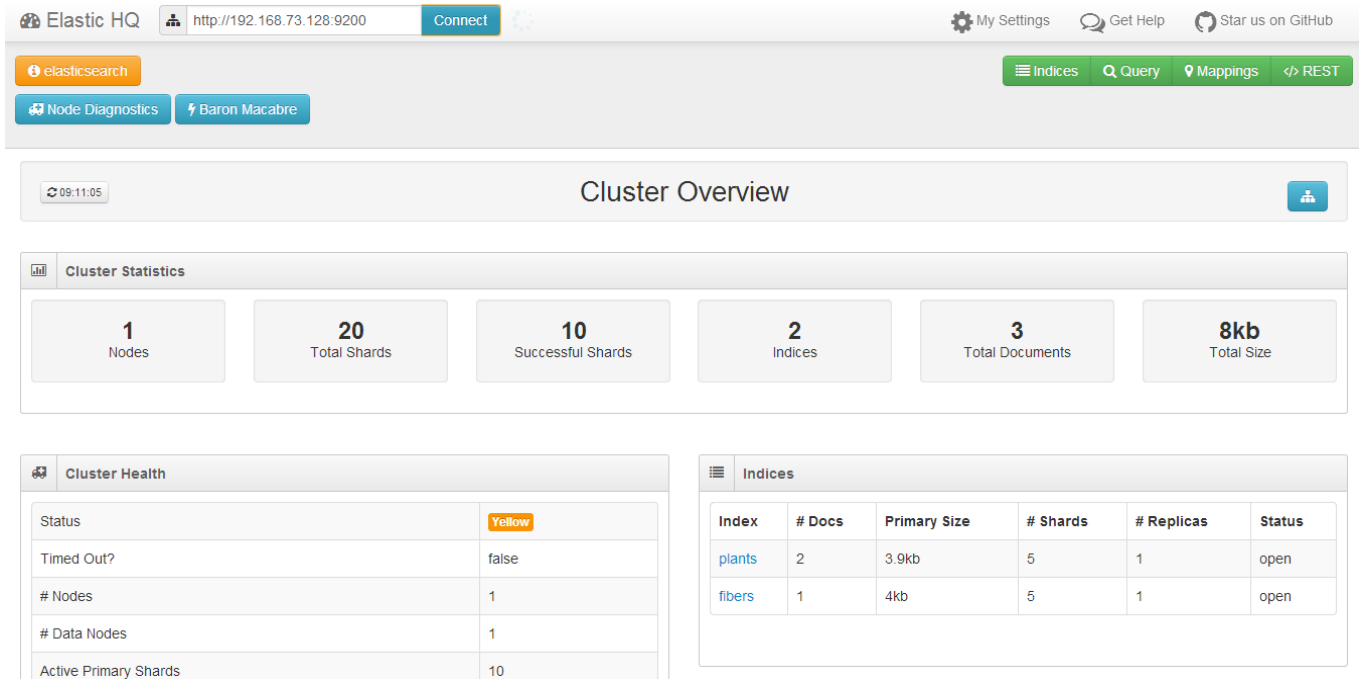
Plugins: Head



Plugins: Kopf



Plugins : ElasticHQ



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

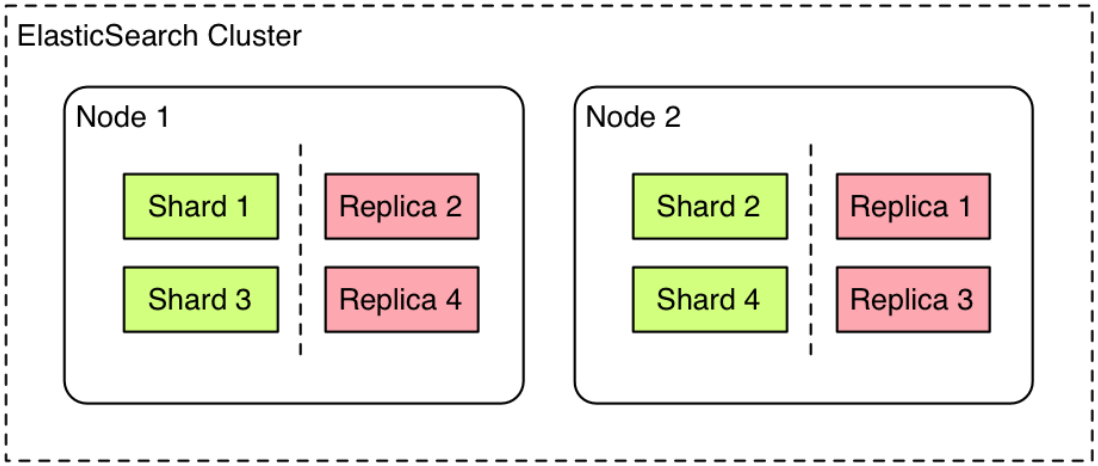
A node is composed by a fixed number of shards

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



Business & Decision

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
}
```

Business & Decision

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
  }
}
```

Business & Decision

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

Analyzers

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"
           }
        }
     }
  }
}
```

Business & Decision

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
          }
        }
      }
    }
  }
}
```

Business & Decision

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

Business & Decision

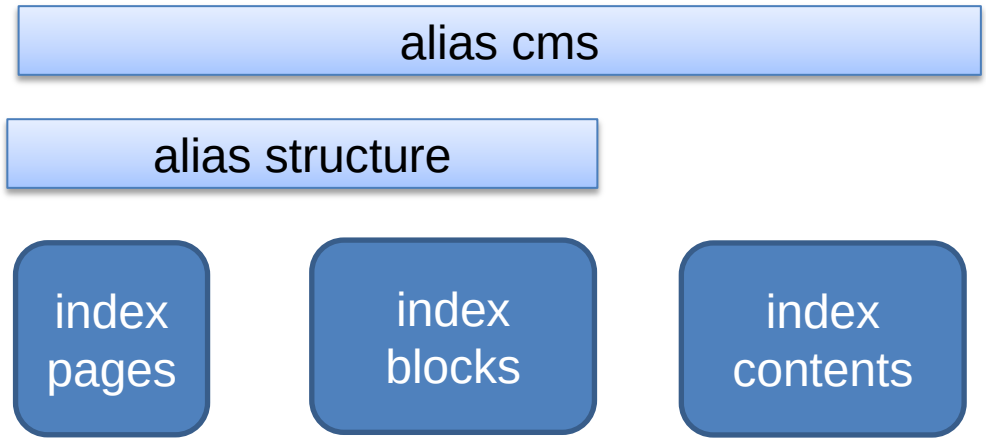
Production

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

Business & Decision

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 ?