

# ELASTICSEARCH



實戰介紹

Kang-min Liu <[gugod@gugod.org](mailto:gugod@gugod.org)>

# ELASTICSEARCH

- Distributed (Near) Real Time Search Engine
- RESTful 風, Lucene 骨, NoSQL 系
- [elasticsearch.org](https://github.com/elasticsearch/elasticsearch)
  - <https://github.com/elasticsearch/elasticsearch>
- [elasticsearch.com](http://elasticsearch.com)

# 名詞對照

Relational DB	ElasticSearch
database	index
table	type
row	document
column	field
schema	mapping
index	(全部)
SQL	query DSL

# CREATE

```
curl -XPOST http://localhost:9200/social/tweet/1 -d '{  
    content: "大家好"  
    user_name: "gugod"  
}'
```

# CREATE

```
curl -XPOST http://localhost:9200/social/tweet/1 -d '{  
    content: "大家好"  
    user_name: "gugod"  
}'
```

# CREATE

```
curl -XPOST http://localhost:9200/social/tweet/1 -d '{  
    content: "大家好"  
  
    user: {  
        name: "gugod",  
        id: 385782393,  
    },  
    tag: ["osdctw", "demo", "moedict"]  
}'
```

# CREATE

```
curl -XPOST http://localhost:9200/social/tweet/1 -d '{  
    content: "大家好"  
    user_name: "gugod"  
}'
```

type

index

# READ

```
curl -XGET http://localhost:9200/social/tweet/1
```

# UPDATE

```
curl -XPUT http://localhost:9200/social/tweet/1 -d '{  
    content: "大家好"  
    user_name: "gugod"  
}'
```

# DELETE

```
curl -XDELETE http://localhost:9200/social/tweet/1
```

# SEARCH

```
$ curl -XGET 'http://localhost:9200/twitter/tweet/_search' -d '{  
  "query": {  
    "filtered": {  
      "query": {  
        "query_string": {  
          "query": "some query string here"  
        }  
      },  
      "filter": {  
        "term": { "user": "kimchy" }  
      }  
    }  
  }  
}'
```

query DSL

# SEARCH

```
$ curl -XPOST 'http://localhost:9200/twitter/tweet/_search' -d '{  
  "query": { "term" : { "user" : "kimchy" } }  
}'
```

# SEARCH

```
$ curl -XGET 'http://localhost:9200/twitter/tweet/_search?q=nihao'
```

# 搜尋引擎原理

- Inverted index 反向索引
  - term → id
  - Relevance Scoring 分數

# 輸入文件

延展性：物質具延長及展開的性質，稱為「延展性」。為大多數金屬之特性。

延平郡王：明鄭成功的封號。

延年益壽：延長壽命，多為頌祝人長壽的用詞。

延性：物質可延長為細絲的性質，稱為「延性」。材料在破壞之前，呈現塑性變形的程度。延性可用拉力試驗中的伸長率及斷面縮率表示之。

延緩：延遲、延後。

# TOKENIZATION

延展性物質具延長及展開的性質稱為延展性為大多數金屬之特性

延平郡王明鄭成功的封號

延年益壽延長壽命多為頌祝人長壽的用詞

延性物質可延長為細絲的性質稱為延性材料在破壞之前呈現塑性變形的程度延性可用拉力試驗中的伸長率及斷面縮率表示之

延緩延遲延後

# TOKENIZATION

延展性：物質 質具 具延 延長 長及 及展 展開 開的 的性 性質 稱為 延展 展性 為大多 多數 數金 金屬 屬之 之特 特性

延平郡王：明鄭 鄭成 成功 功的 的封 封號

延年益壽：延長 長壽 壽命 多為 為頌 頌祝 祝人 人長 長壽 壽的 的用 用詞

延性：物質 質可 可延 延長 長為 為細 細絲 絲的 的性 性質 稱為 延性 材料 料在 在破 破壞 壞之 之前 呈現 現塑 塑性 性變 變形 形的 的程 程度 延性 性可 可用 用拉 拉力 力試 試驗 驗中 中的 的伸 伸長 長率 率及 及斷 斷面 面縮 縮率 率表 表示 示之

延緩：延遲 延後

# 搜尋：延長

**延展性：**物質 質具 具延 延長 長及 及展 展開 開的 的性 性質 稱為 延展 展性 為大多 多數 數金 金屬 屬之 之特 特性

延平郡王：明鄭 鄭成 成功 功的 的封 封號

**延年益壽：**延長 長壽 壽命 多為 為頌 頌祝 祝人 人長 長壽 壽的 的用 用詞

**延性：**物質 質可 可延 延長 長為 為細 細絲 絲的 的性 性質 稱為 延性 材料 料在 在破 破壞 壞之 之前 呈現 現塑 塑性 性變 變形 形的 的程 程度 延性 性可 可用 用拉 拉力 力試 試驗 驗中 中的 的伸 伸長 長率 率及 及斷 斷面 面縮 縮率 率表 表示 示之

延緩：延遲 延後

# 給分

**延展性：**物質 質具 具延 延長 長及 及展 展開 開的 的性 性質 稱為 延展 展性 為大多 多屬 金屬 屬之 之特 特性

0.5

延平郡王：明鄭 鄭成 功的 的封 封號

**延年益壽：**延長 長為 為頌 頌祝 祝人 人長 長壽 壽的 的用 用詞

**延性：**物質 質可 可延 延長 長為 為細 細絲 絲的 的性 性質 稱為 延性 材料 在 在破 破壞 伸 呈現 現塑 塑性 性變 變形 形的 的程 程度 延性 性可 可用 用拉 拉力 力試 試驗 驗中 中的 的伸 伸長 長率 率及 及斷 斷面 面縮 縮率 率表 表示 示之

0.3

延緩：延遲 延後

# 給分

- 文件的詞數（長度）
- 詞在文件內的頻率
- 詞在索引內的頻率
- 任意指定

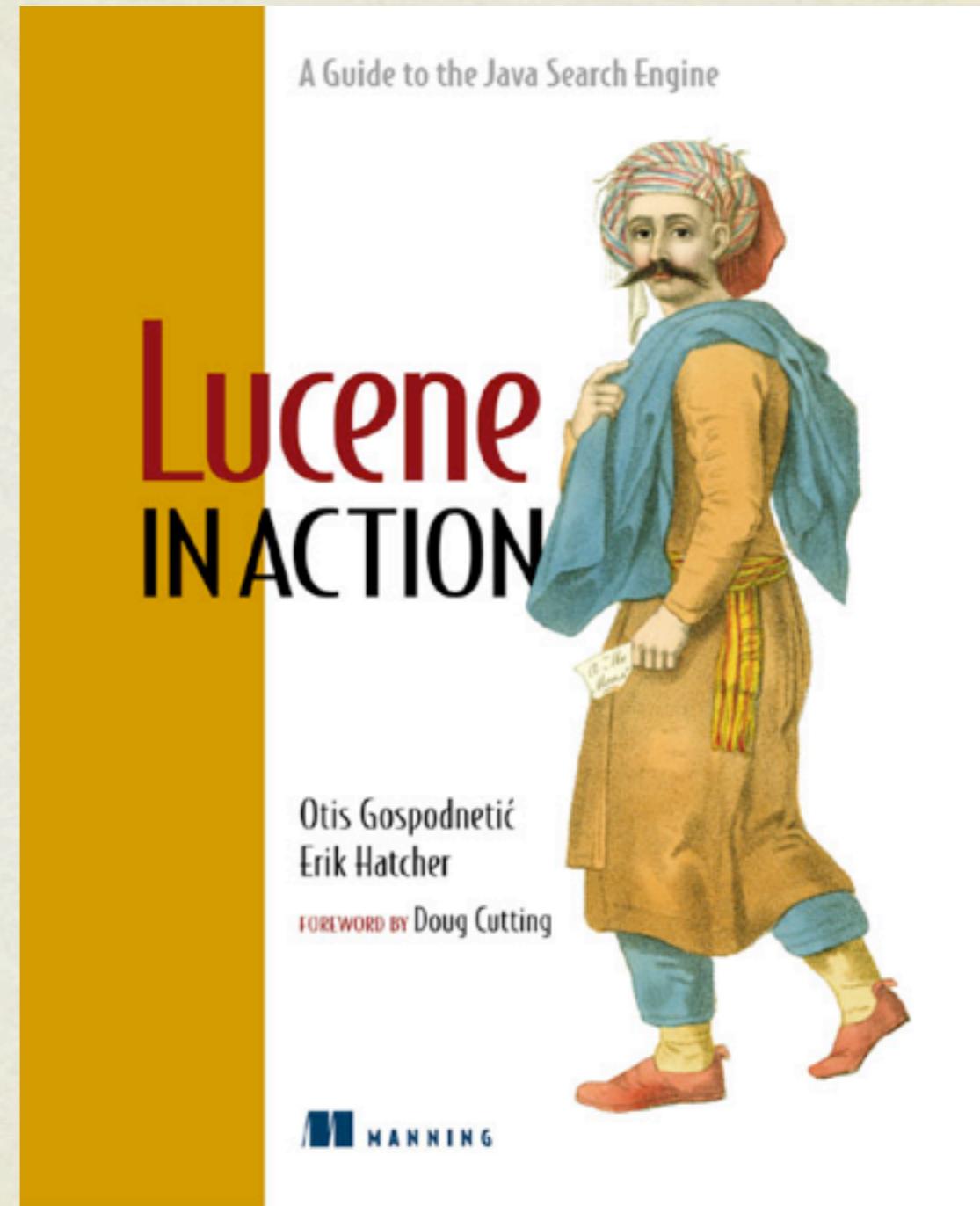
# 給分

- 文件的詞數（長度）
- 詞在文件內的頻率
- 詞在索引內的頻率
- 任意指定

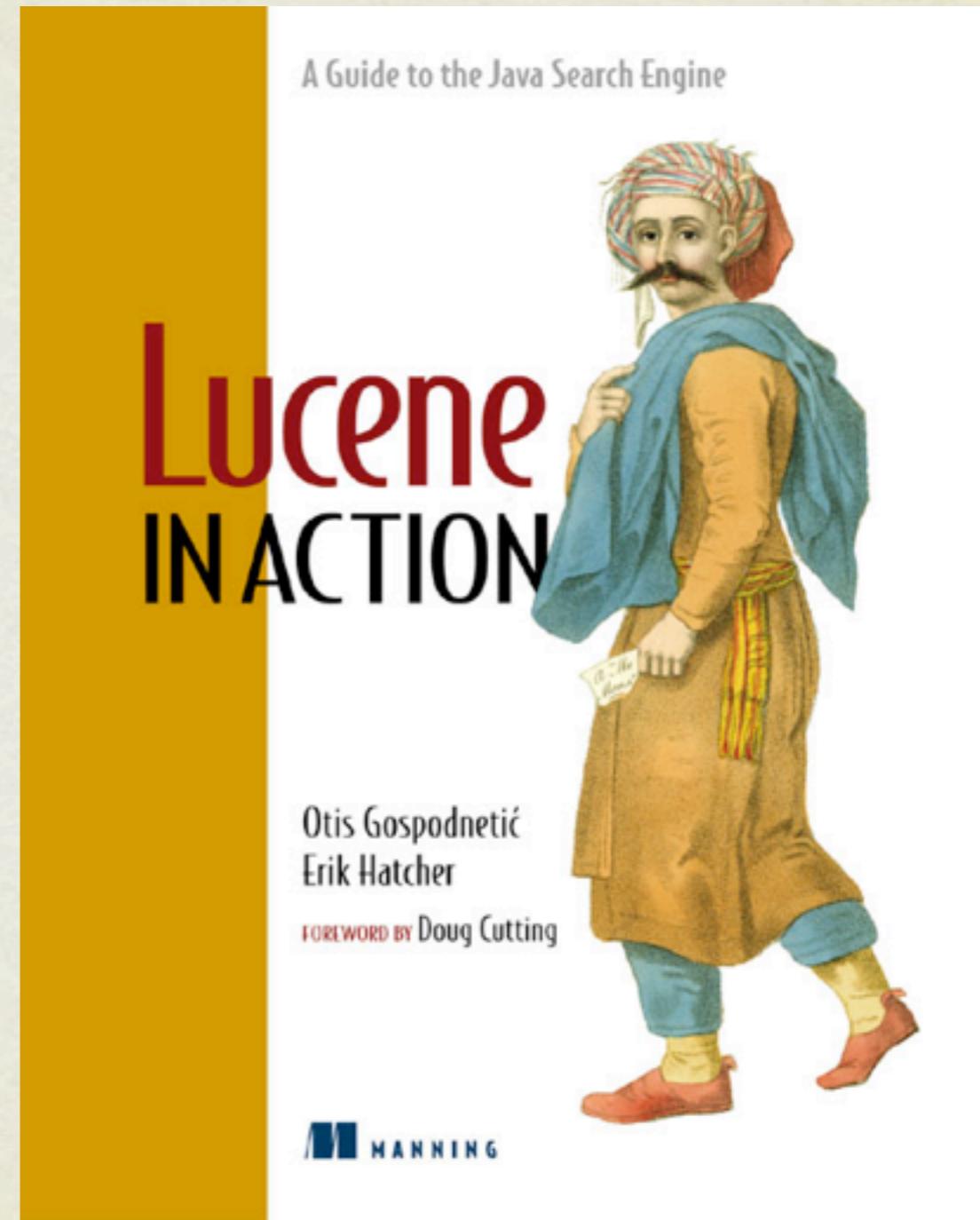
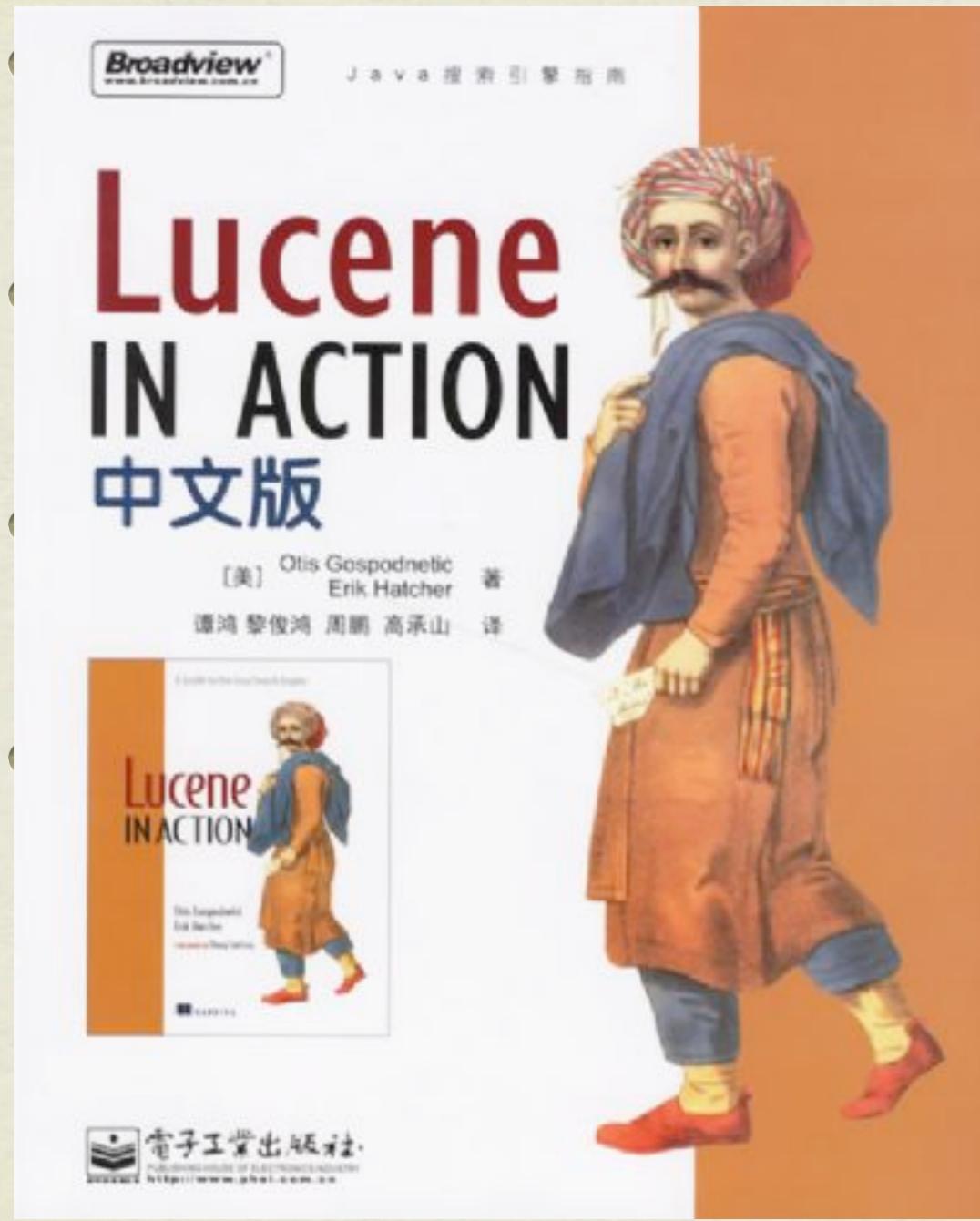


# 給分

- 文件的詞數（長度）
- 詞在文件內的頻率
- 詞在索引內的頻率
- 任意指定



# 給分



# QUERY

# TERM

```
{  
  "query": {  
    "term": {  
      "user": "ingy"  
    }  
  }  
}
```

# TEXT

```
{  
  "query": {  
    "text": {  
      "content": "這樣那樣"  
    }  
  }  
}
```

# TEXT

```
{  
  "query": {  
    "text": {  
      "content": "這樣那樣"  
    }  
  }  
}
```

這樣 樣那 那樣

# RANGE

```
{  
  "query": {  
    "range" : {  
      "age" : {  
        "from" : 10,  
        "to" : 20,  
        "include_lower" : true,  
        "include_upper": false,  
        "boost" : 2.0  
      }  
    }  
  }  
}
```

# QUERY\_STRING

```
{  
  "query": {  
    "query_string" : {  
      "query" : "這樣 AND 那樣 OR 怎樣"  
    }  
  }  
}
```

# WILDCARD

```
{  
  "query": {  
    "wildcard" : { "user" : "ki*y" }  
  }  
}
```

# MLT(MORE LIKE THIS)

```
{  
  "query": {  
    "more_like_this": {  
      "like_text": "這樣那樣",  
      "min_term_freq": 1,  
      "max_query_terms": 12  
    }  
  }  
}
```

# 萌典 + ES

<https://github.com/g0v/esmoe>

# 文件結構

```
{  
  "heteronyms": [  
    {  
      "bopomofo": "一ㄢˊ ㄓㄢˇ ㄉㄧㄥˇ",  
      "bopomofo2": "yán jiǎn shèng",  
      "definitions": [  
        {  
          "def": "物質具延長及展開的性質，稱為「延展性」。為大多  
數金屬之特性。"  
        }  
      ],  
      "pinyin": "yán zhǎn xìng"  
    },  
    ],  
    "title": "延展性"  
  },
```

# 正查

```
# curl http://localhost:9200/moedict/revised/$(uri_escape 搜)?pretty=1
{
  "_index": "moedict",
  "_type": "revised",
  "_id": "搜",
  "_version": 1,
  "exists": true, "_source": {"non_radical_stroke_count":10,"stroke_count":13,"heteronyms":[{"bopomofo":"厃又","pinyin":"sōu","bopomofo2":"sōu","definitions":[{"quote":["漢書。卷六。武帝紀：「秋，閉城門大搜，發謫戍屯五原。」","聊齋志異。卷一。狐嫁女：「已而主人斂酒具，少一爵，冥搜不得。」"],"def":"找尋、尋求。","type":"動"}],"example":[{"如：「搜身」。"}, {"quote":["元。王實甫。西廂記。第三本。第二折：「不肯搜自己狂為，只待要覓別人破綻。」"],"def":"檢查、檢點。","type":"動"}]}],"title": "搜", "radical": "手"}
```

# 反查（由義查詞）

```
# curl --silent http://localhost:9200/moedict/revised/_search\  
'?fields=&pretty=1&q='$(uri_escape 找尋) | grep _id  
  
"_id" : "尋根",  
"_id" : "尋找",  
"_id" : "訪求",  
"_id" : "探礦",  
"_id" : "尋求",  
"_id" : "找尋",  
"_id" : "找機會",  
"_id" : "自找",  
"_id" : "尋樂",  
"_id" : "探索",
```

# 用例句查

```
# curl --silent -XPOST http://localhost:9200/moedict/revised/_search?pretty'\n-d '{query:{text:{example: "紅樓夢" }}}'\n{\n    \"took\" : 8,\n    \"timed_out\" : false,\n    \"_shards\" : {\n        \"total\" : 5,\n        \"successful\" : 5,\n        \"failed\" : 0\n    },\n    \"hits\" : {\n        \"total\" : 485,\n        \"max_score\" : 5.116848,\n        \"hits\" : [ {\n            \"_index\" : \"moedict\",\n            \"_type\" : \"revised\",\n            \"_id\" : \"抄本\",\n            \"_score\" : 5.116848, \"_source\" : {\"heteronyms\" :[ {\"bopomofo\" : \"ㄔㄠ ㄅㄣ\", \"pinyin\" : \"chāo běn\", \"bopomofo2\" : \"chāu běn\", \"definitions\" : [ {\"link\" : [ \"亦稱為「寫本」、「鈔本」。\" ], \"example\" : [ \"如：「抄本紅樓夢」。\" ], \"synonyms\" : \"手本\", \"def\" : \"手抄的書籍。\" } ] }, \"title\" : \"抄本\" }\n        }, {\n            \"_index\" : \"moedict\",\n            \"_type\" : \"revised\",\n            \"_id\" : \"一名\",\n            \"_score\" : 4.27241, \"_source\" : {\"heteronyms\" :[ {\"bopomofo\" : \"一 ㄇㄧㄥˊ\", \"pinyin\" : \"yī míng\", \"bopomofo2\" : \"vī míng\", \"definitions\" : [ {\"quote\" : \"紅樓夢 第四十八回：「派下薛蟠之乳父\" } ] }\n        }\n    }\n}
```

# 用引言查

```
# curl --silent -XPOST http://localhost:9200/moedict/revised/_search?pretty'\n-d '{query:{text:{quote: "紅樓夢"}}}'\n{\n    \"took\" : 18,\n    \"timed_out\" : false,\n    \"_shards\" : {\n        \"total\" : 5,\n        \"successful\" : 5,\n        \"failed\" : 0\n    },\n    \"hits\" : {\n        \"total\" : 11858,\n        \"max_score\" : 1.7719736,\n        \"hits\" : [ {\n            \"_index\" : \"moedict\",\n            \"_type\" : \"revised\",\n            \"_id\" : \"原稿\",\n            \"_score\" : 1.7719736, \"_source\" : {"heteronyms": [{"bopomofo": "ㄩㄢˊ", "pinyin": "yuán gǎo", "bopomofo2": "yuán gǎu"}, "definitions": [{"synonyms": "底稿, 稿本, 初稿, 草稿", "quote": ["紅樓夢。第五回：「說畢，回頭命小鬟取了『紅樓夢』原稿來，遞與寶玉。」"], "def": "作品最初的手稿。"}]}, "title": \"原稿\"}, {\n            \"_index\" : \"moedict\",\n            \"_type\" : \"revised\",\n            \"_id\" : \"一面之緣\",
```

# 用注音查

# 出現頻率最高的注音

```
# curl --silent -XPOST http://localhost:9200/moedict/revised/_search?pretty'  
-d '  
{  
  "query" : {  
    "match_all" : {}  
  },  
  "facets" : {  
    "bpmf" : {  
      "terms" : {  
        "field" : "bopomofo"  
      }  
    }  
  }  
}'
```

```
{  
  "facets" : {  
    "bpmf" : {  
      "_type" : "terms",  
      "missing" : 1485,  
      "total" : 430736,  
      "other" : 401706,  
      "terms" : [ {  
        "term" : "\u201e\u201c",  
        "count" : 5400  
      }, {  
        "term" : "\u201e",  
        "count" : 4438  
      }, {  
        "term" : "\u201c",  
        "count" : 3218  
      }, {  
        "term" : "\u2014",  
        "count" : 2939  
      }, {  
        "term" : "\u201e\u201c",  
        "count" : 2754  
      }, {  
        "term" : "\u201e",  
        "count" : 2625  
      }, {  
        "term" : "\u201c",  
        "count" : 2580  
      }, {  
        "term" : "\u201d",  
        "count" : 2044  
      }, {  
        "term" : "\u201e\u201d",  
        "count" : 1541  
      }, {  
        "term" : "\u201d\u201c",  
        "count" : 1491  
      } ]  
    } ]  
}
```

# FACETS

# FACETS

- Aggregations
- `SELECT SUM(salary) GROUP BY name FROM employee;`

# FACETS

- range
- term stats
- geo distance
- statistical
- date histogram

# 例：立院公報

{

"speaker" : "陳委員清寶",

"content" : "我在推動小三通的時候，很多人都沒有注意到我真正的用意，除了是為當地爭取繁榮進步、增加建設的機會之外，我最著眼希望的是藉由小三通來催化全面的三通，能夠讓兩岸關係解凍。從這一個點切入，找一些議題與大陸進行對談。",

"issue" : "3109"

}

{

"speaker" : "李委員慶華",

"content" : "希望包括核一、核二、核三、核四的安全係數都能提高。",

"issue" : "3867"

}

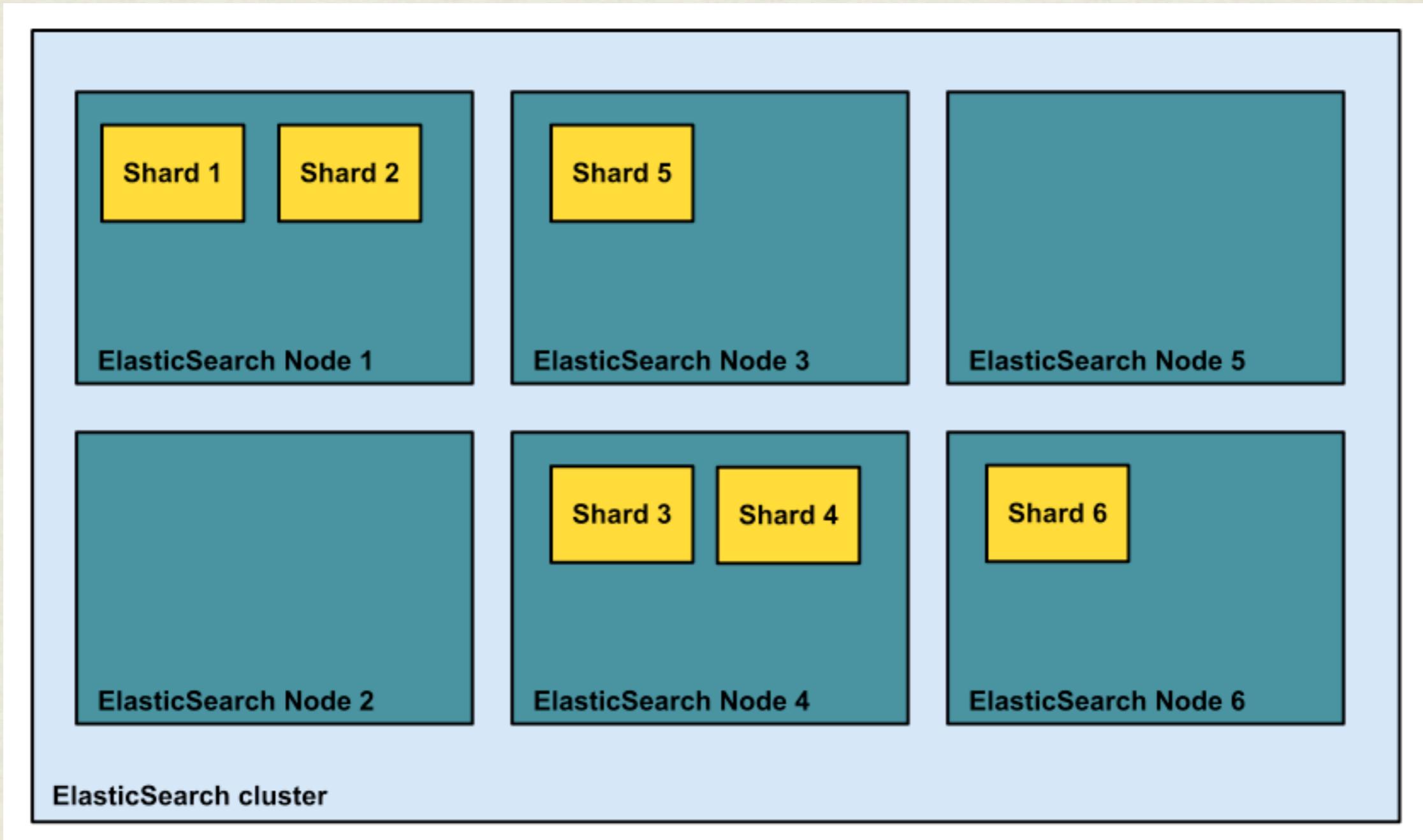
```
curl -XPOST http://localhost:9200/_search?pretty=1' -d '  
{  
  "query": {  
    "match_all": {}  
  },  
  "facets": {  
    "top": {  
      "terms": {  
        "field": "content"  
      }  
    }  
  }'  
}'
```

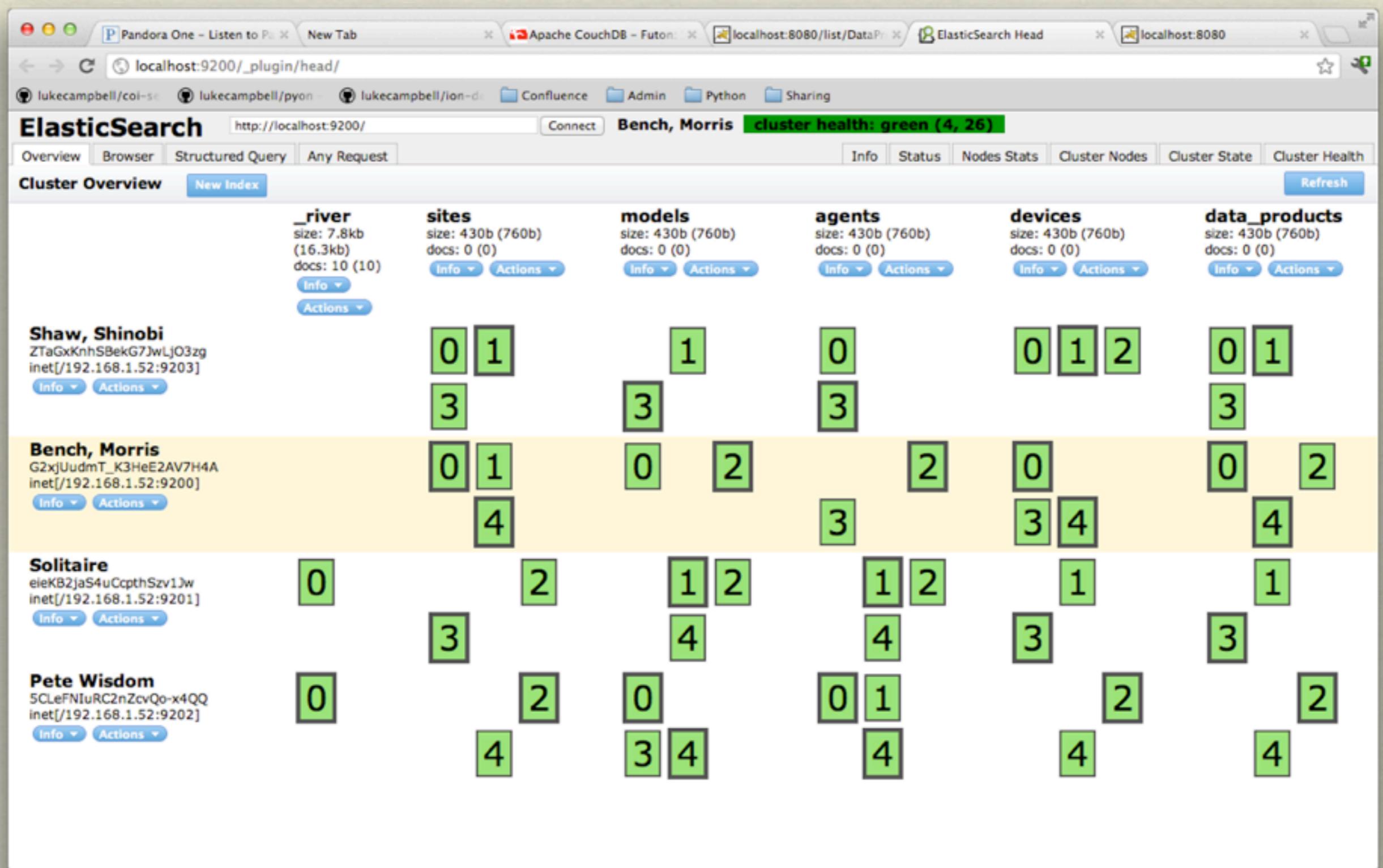
```
curl -XPOST http://localhost:9200/_search?pretty=1' -d '  
{  
  "query": {  
    "match_all": {}  
  },  
  "facets": {  
    "top": {  
      "terms": {  
        "field": "speaker"  
      }  
    }  
  }'  
}'
```

# OPERATION

# CLUSTERING

- auto-discovery
- auto-elected master
- data replication / partition
  - with flexible shard / replica setting





- shard ~ partition
- replica ~ duplication

# CLUSTERING

- more shard
  - faster indexing / scaling
- more replica
  - faster searching / failover

# ElasticSearch

http://192.168.7.8:9200/

Connect

Rick

cluster health: yellow (6, 18)

Overview

Browser

Structured Query

Any Request

Info

Status

Nodes Stats

Cluster Nodes

Cluster State

Cluster Health

Cluster Overview

New Index

cu\_docs

size: 180Gb (540Gb)  
docs: 995131 (995131)

Info Actions

bnil

size: 80kb (480kb)  
docs: 90 (90)

Info Actions

cu\_msg

size: 313Gb (1.56Tb)  
docs: 10047450 (10140915)

Info Actions

anvil

index: close

Info Actions

Leon 3Wqr1xaCRu-b0uEzDkmrDg  
inet[/192.168.7.8:9202]

Info Actions

0 1

0 1

0

Info Actions

Pris L8qx7ilfSI-kcKq\_6bMbWw  
inet[/192.168.7.8:9204]

Info Actions

0 1

0 1

0

Info Actions

Rick Vnpri1FNTGirwRfZsZ2RxQ  
inet[/192.168.7.8:9200]

Info Actions

1 2

0 1

0 1 2 3 4

Info Actions

Rachel 87KsIv0FTVSkkqwENaja6A  
inet[/192.168.7.8:9203]

Info Actions

1 2

0 1

0 1 2 3 4

Info Actions

Zhora b6NxRTxsR\_WUQl5cXPKhbw  
inet[/192.168.7.8:9205]

Info Actions

0 2

0 1

0 1 2 3 4

Info Actions

Roy \_8Rj2wYVT7Svn\_v5F97jJA  
inet[/192.168.7.8:9201]

Info Actions

0 2

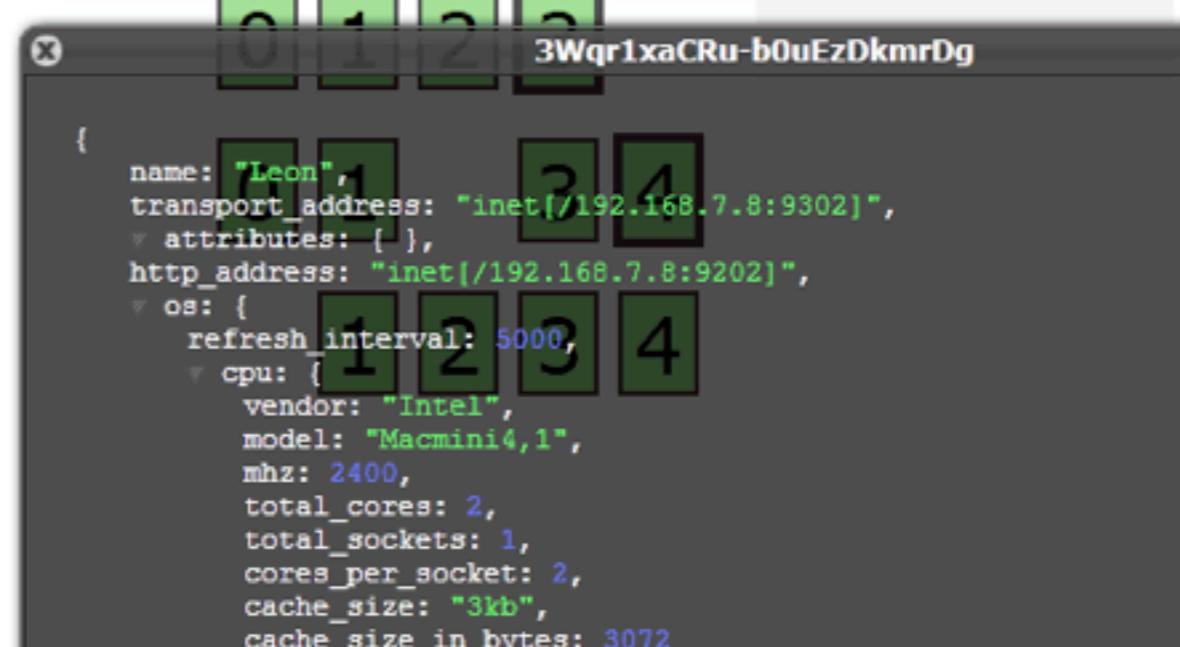
0 1

0 1 2 3 4

Info Actions

Unassigned

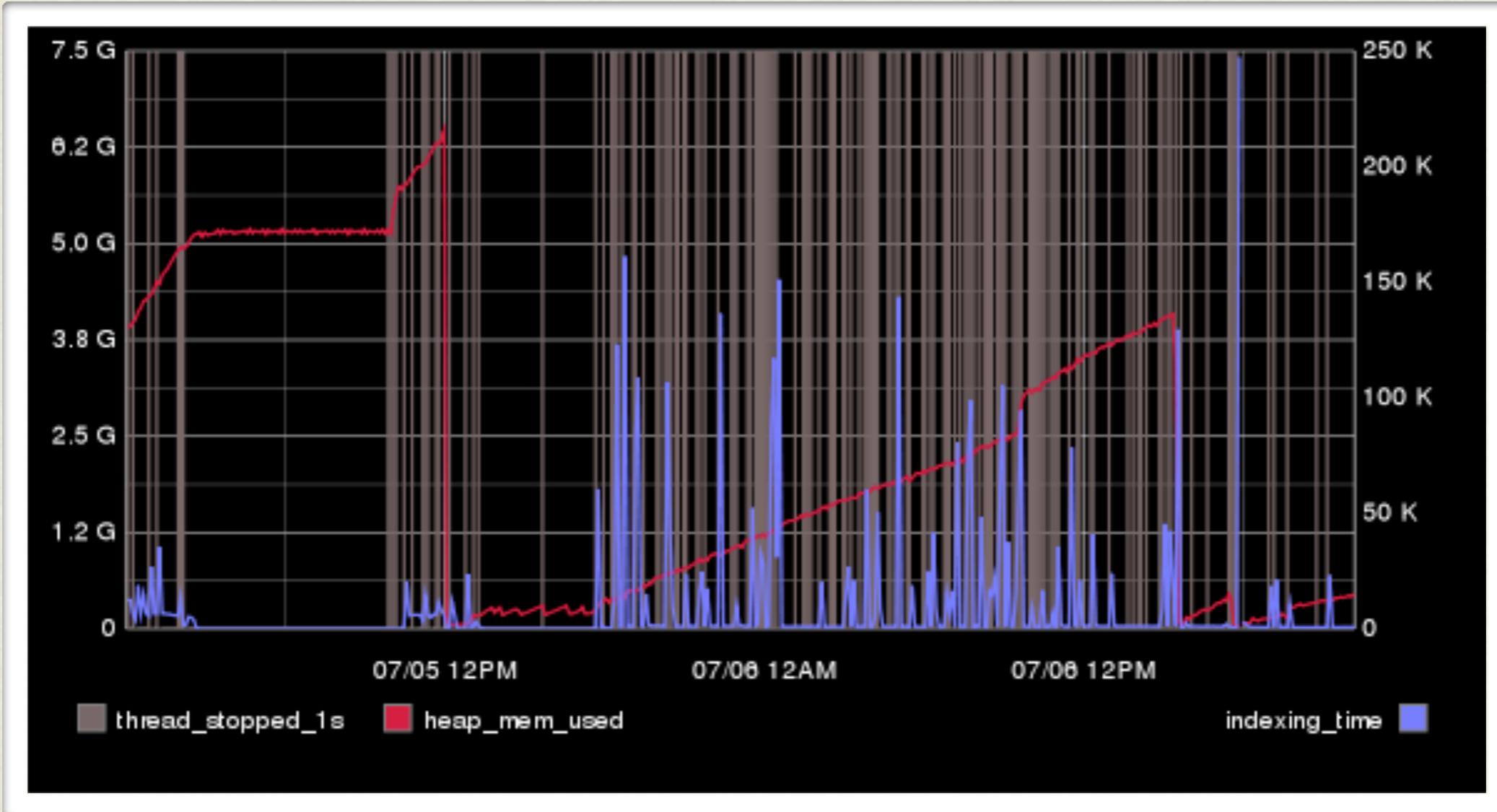
0 0  
1 1



# STATS API

- <http://www.elasticsearch.org/guide/reference/api/admin-cluster-nodes-stats/>
- 取得各項數據
- 文件數、搜尋次數、累計搜尋時數、累計建索引時間
  - cluster / primary / node / index 各種級別
- JVM CPU/Heap / OS / Thread / transport 使用狀態

# JVM GC 圖



# SEE ALSO

- Cool, Bonsai Cool - An introduction to ElasticSearch  
<http://bit.ly/112xtsk>
- The Road to a Distributed Search Engine  
<http://bit.ly/ZqBBUt>
- elasticsearch, Big Data, Search & Analytics  
<http://bit.ly/11tmbyK>



—END—  
*OKTHXBYE*