

# OPEN SOURCE SEARCH ENGINE

---

**Planning Team**

**2020/04**

# AGENDA

- 🎯 Introduction
- 🎯 Effective Search Engine
- 🎯 Open Source Search Engine
- 🎯 Comparison: Apache Solr vs. Elasticsearch vs. Sphinx



# INTRODUCTION

- ❑ **Search engine (platform)** là một phần mềm dùng để tìm kiếm dữ liệu dựa trên các thông tin cụ thể.
- ❑ Có 3 loại search engine phổ biến:
  - **Web / Internet search:** Google, Bing, Yahoo!, Baidu...
  - **Enterprise search:** Apache Solr, Elasticsearch...
  - **Desktop search:** Windows search, Spotlight (Mac OS)...



# AGENDA

- 🎯 Introduction
- 🎯 **Effective Search Engine**
- 🎯 Open Source Search Engine
- 🎯 Comparison: Apache Solr vs. Elasticsearch vs. Sphinx



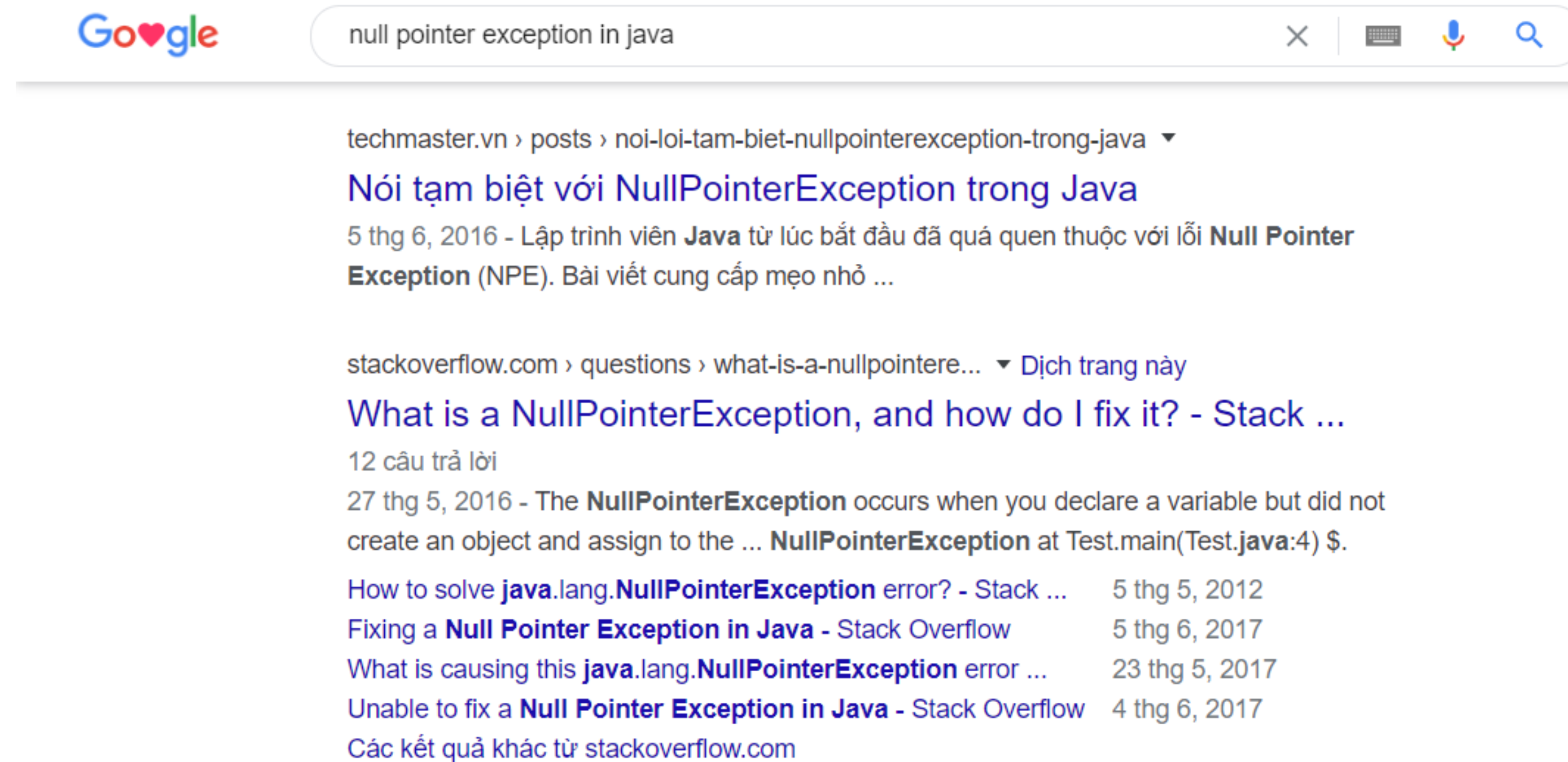
# **EFFECTIVE SEARCH ENGINE**

- ❑ **Search engine** được đánh giá là **hiệu quả (effective)** khi có thể tìm kiếm những kết quả **liên quan nhất có thể** so với nội dung tìm kiếm.
- ❑ Một số **tính năng, kỹ thuật tìm kiếm** mà một search engine hiệu quả nên có:
  - ✓ Full-text search
  - ✓ Highlighting
  - ✓ Faceted search
  - ✓ Fuzzy search
  - ✓ Geospatial search



# TECHNIQUE: FULL-TEXT SEARCH

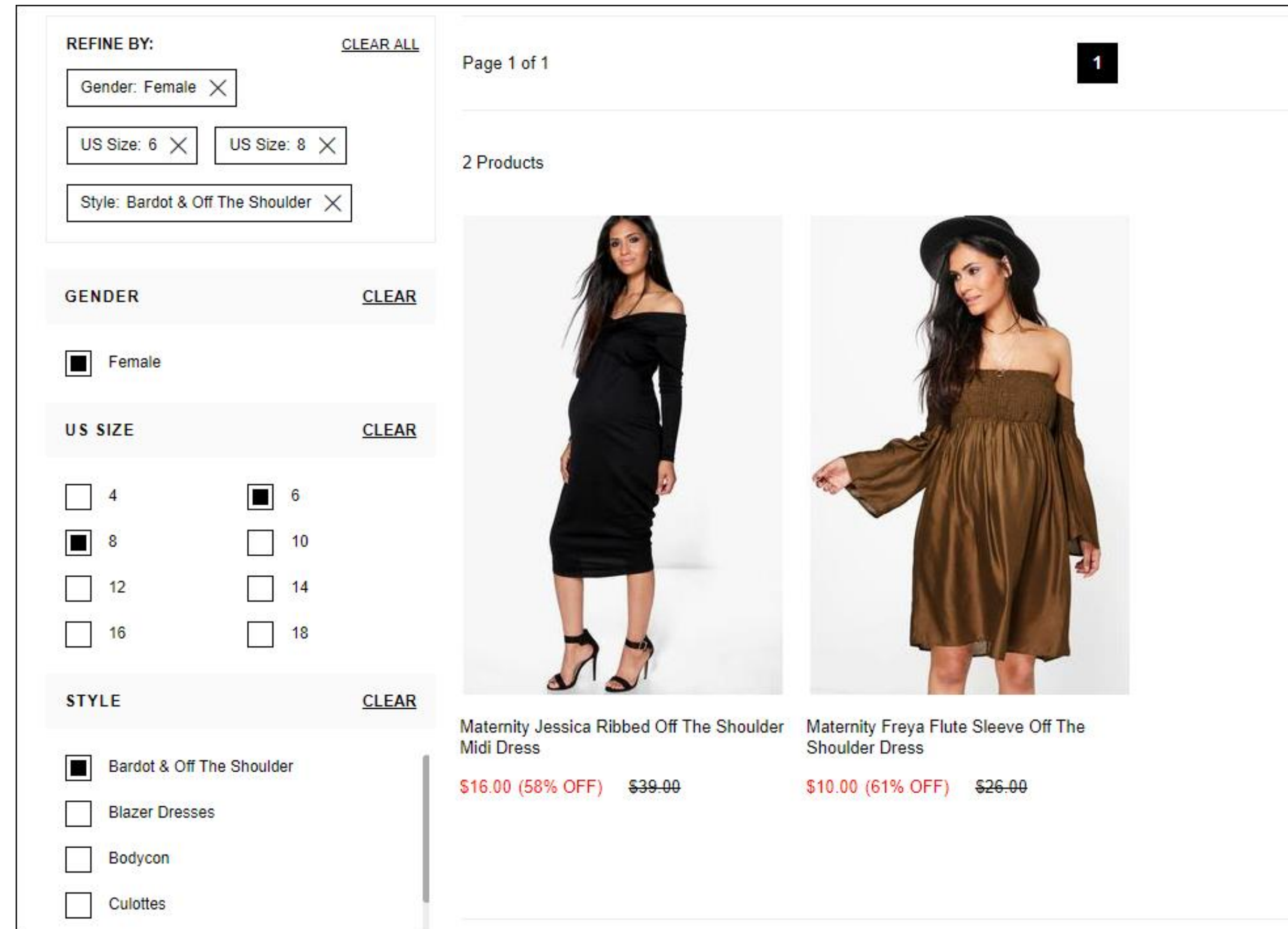
**Full-text search** là một kỹ thuật tìm kiếm dữ liệu được tạo ra nhằm **thay thế cho câu query LIKE** khi cần tìm kiếm dữ liệu **linh hoạt hơn, hiệu suất tốt hơn, khắc phục được những giới hạn** mà câu query **LIKE** gặp phải.





# TECHNIQUE: FACETED SEARCH

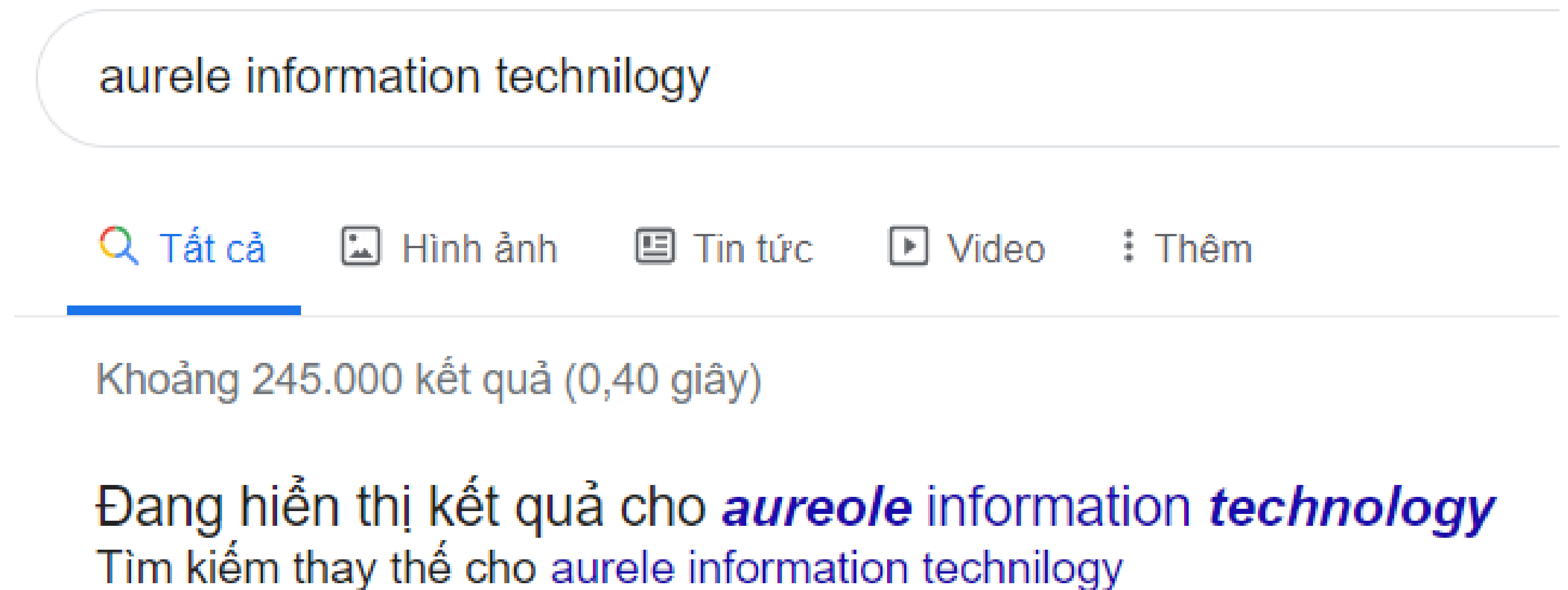
Faceted search là kỹ thuật tìm kiếm dữ liệu dựa trên các **keyword** (nếu có) và **nhiều filter param kết hợp lại với nhau** để trả về kết quả tìm kiếm liên quan nhất có thể.



# TECHNIQUE: FUZZY SEARCH

**Fuzzy search** là kỹ thuật tìm kiếm gần đúng (**Approximate search**) dựa trên nền tảng thuật toán **Levenshtein distance**, tìm kiếm theo **3 phép biến đổi keyword**:

- ✓ Thêm 1 ký tự
- ✓ Bớt 1 ký tự
- ✓ Thay đổi 1 ký tự



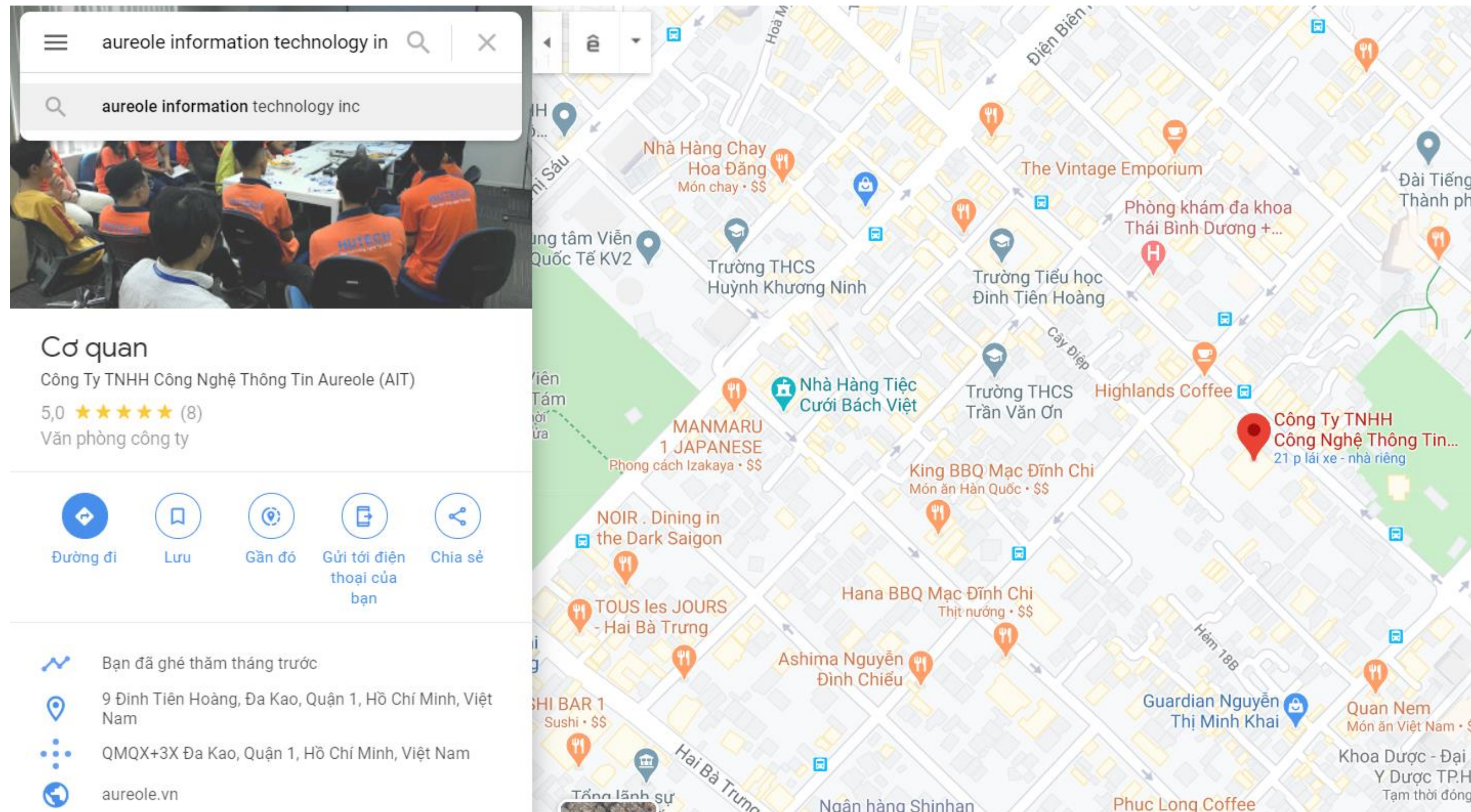
The screenshot shows a Google search bar with the text "aurele information technilogy". Below the search bar, there are filters for "Tất cả", "Hình ảnh", "Tin tức", "Video", and "Thêm". Below the filters, it says "Khoảng 245.000 kết quả (0,40 giây)". At the bottom, it says "Đang hiển thị kết quả cho aureole information technology" and "Tìm kiếm thay thế cho aurele information technology".





# TECHNIQUE: GEOSPATIAL SEARCH

**Geospatial search** là kỹ thuật tìm kiếm vị trí theo vĩ độ (latitude) và kinh độ (longitude).





# AGENDA

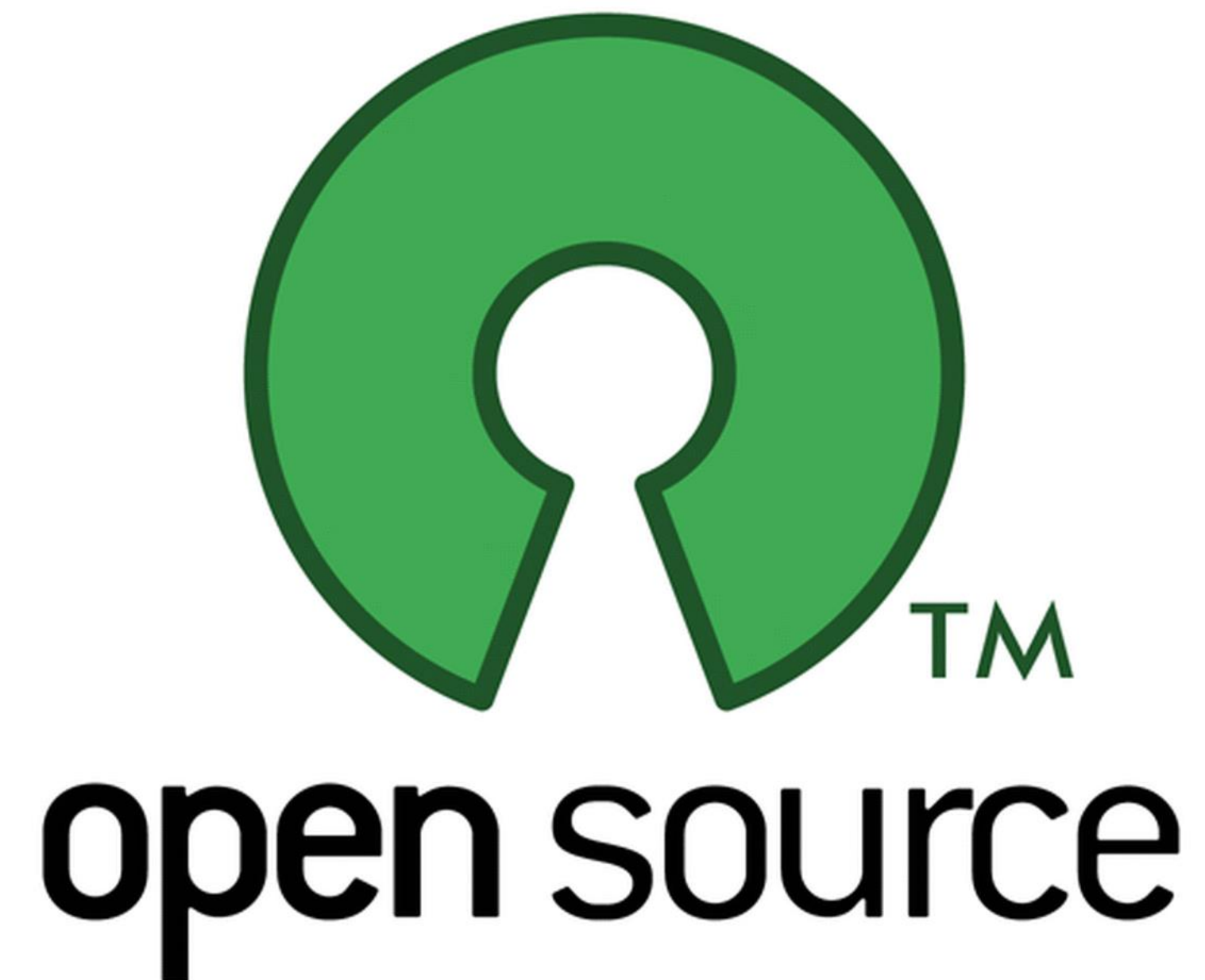
- 🎯 Introduction
- 🎯 Effective Search Engine
- 🎯 **Open Source Search Engine**
- 🎯 Comparison: Apache Solr vs. Elasticsearch vs. Sphinx



# **OPEN SOURCE SEARCH ENGINE**

Một số open source search engine:

- 🔍 Apache Lucene
- 🔍 Apache Solr
- 🔍 Elasticsearch
- 🔍 Sphinx
- 🔍 Indica
- 🔍 Splunk



# OPEN SOURCE SEARCH ENGINE: RANKING

☐ include secondary database models

21 systems in ranking, April 2020

Rank			DBMS	Database Model	Score		
Apr 2020	Mar 2020	Apr 2019			Apr 2020	Mar 2020	Apr 2019
1.	1.	1.	Elasticsearch	Search engine, Multi-model	148.91	-0.26	+2.91
2.	2.	2.	Splunk	Search engine	88.08	-0.44	+4.99
3.	3.	3.	Solr	Search engine	53.59	-1.50	-6.64
4.	4.	4.	MarkLogic	Multi-model	11.26	-0.67	-3.21
5.	5.	5.	Sphinx	Search engine	6.49	-0.20	-0.48
6.	6.	6.	Microsoft Azure Search	Search engine	6.22	-0.12	+0.57
7.	7.	7.	ArangoDB	Multi-model	4.88	-0.06	+0.59
8.	8.	8.	Algolia	Search engine	4.45	-0.17	+0.36
9.	10.	11.	Amazon CloudSearch	Search engine	2.69	-0.16	-0.31
10.	9.	9.	Virtuoso	Multi-model	2.62	-0.24	-0.70
11.	11.	10.	Google Search Appliance	Search engine	2.62	-0.09	-0.46
12.	12.	12.	Xapian	Search engine	0.69	-0.01	-0.08
13.	13.	13.	CrateDB	Multi-model	0.65	+0.03	-0.07
14.	14.	14.	SearchBlox	Search engine	0.29	-0.01	+0.01
15.	15.	17.	searchxml	Multi-model	0.10	+0.00	+0.05
16.	16.	15.	Manticore Search	Search engine	0.08	+0.00	+0.03
17.	17.		Weaviate	Search engine, Multi-model	0.07	+0.01	
18.	19.	16.	DBSight	Search engine	0.04	+0.00	-0.01
19.	18.	20.	Indica	Search engine	0.03	-0.03	+0.03
20.	20.	18.	Exorbyte	Search engine	0.03	+0.02	-0.01
21.	21.	19.	FinchDB	Multi-model	0.02	+0.02	0.00



# AGENDA

- 🎯 Introduction
- 🎯 Effective Search Engine
- 🎯 Open Source Search Engine
- 🎯 **Comparison: Apache Solr vs. Elasticsearch vs. Sphinx**



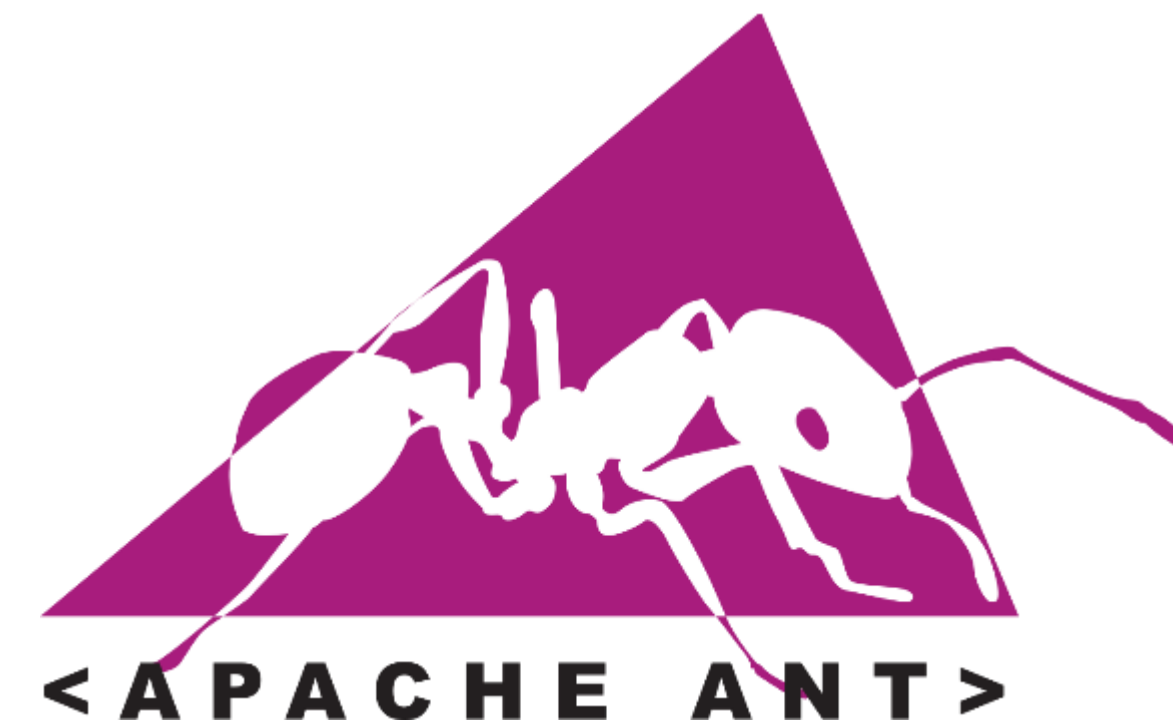
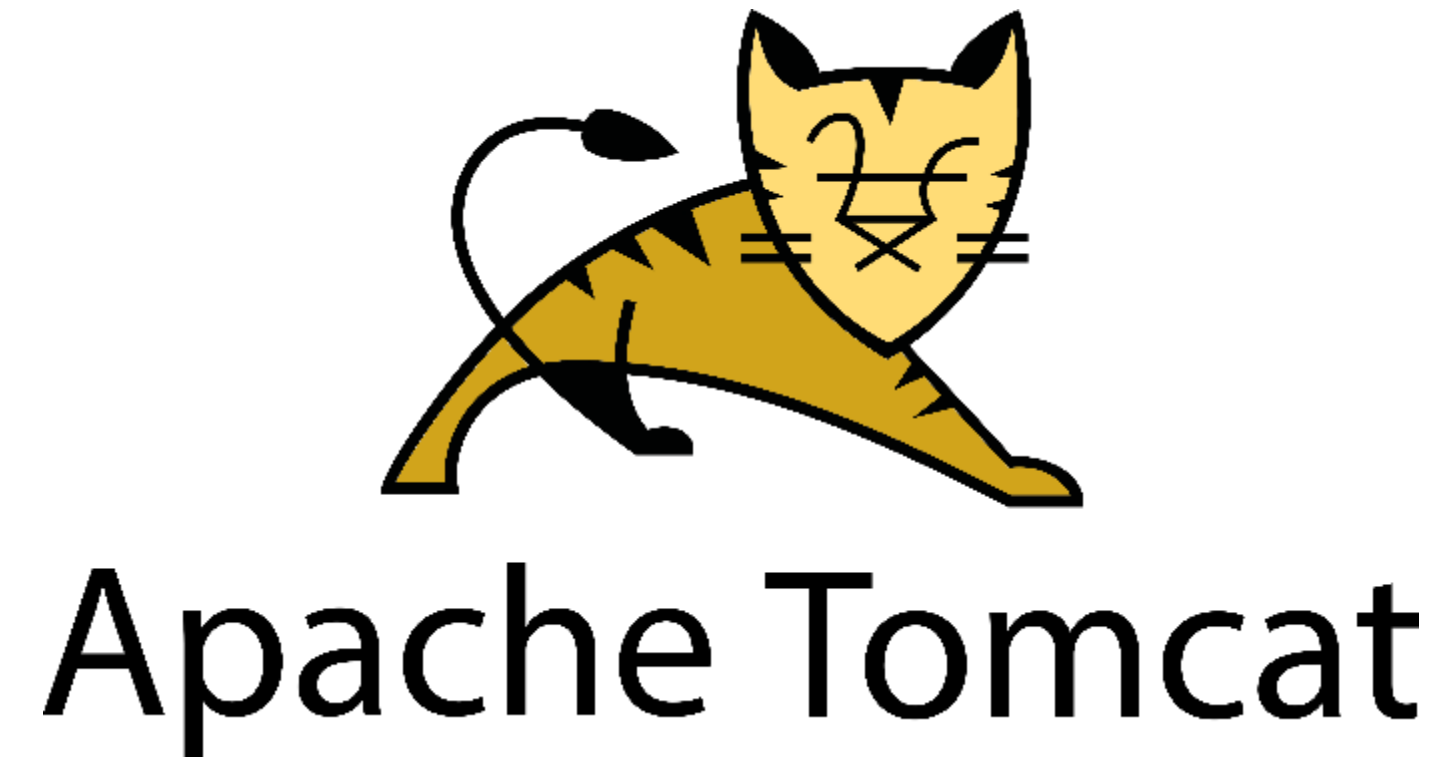
# APACHE SOLR: INTRODUCTION

- ❑ **Apache Solr** là một công cụ tìm kiếm và phân tích được phát triển lại bởi **Apache Software Foundation** (Viết lại từ **Solr** phát triển bởi **CNET Network**), phát hành phiên bản đầu tiên vào năm **2008**.
- ❑ Được phát triển bằng ngôn ngữ **Java**, dựa trên nền tảng thư viện **Apache Lucene**.
- ❑ Website: <https://lucene.apache.org/solr/>





# APACHE SOFTWARE FOUNDATION



# APACHE SOLR: INTRODUCTION

- ❑ **Apache Solr** là một công cụ tìm kiếm và phân tích được phát triển lại bởi **Apache Software Foundation** (Viết lại từ **Solr** phát triển bởi **CNET Network**), phát hành phiên bản đầu tiên vào năm **2008**.
- ❑ Được phát triển bằng ngôn ngữ **Java**, dựa trên nền tảng thư viện **Apache Lucene**.
- ❑ Website: <https://lucene.apache.org/solr/>





## Dashboard

Logging

Cloud

Collections

Java Properties

Thread Dump



Collection Sele...

Core Selector

## Instance

Start about 4 hours ago

## Versions


 solr-spec 6.0.0  
solr-impl 6.0.0 48c80f91b8e5cd9b3a9b48e6184bd53e7619e7e3 -...  
 lucene-spec6.0.0  
lucene-impl6.0.0 48c80f91b8e5cd9b3a9b48e6184bd53e7619e7e3 -...

25.70 MB

## JVM

 Runtime Oracle Corporation Java HotSpot(TM) 64-Bit Server VM 1.8....

 Processors 4

 Args -DSTOP.KEY=solrrocks  
-DSTOP.PORT=7983  
-Djetty.home=/tmp/solr-6.0.0/server  
-Djetty.port=8983  
-Dlog4j.configuration=file:/tmp/solr-6.0.0/example/resourc...  
-Dsolr.install.dir=/tmp/solr-6.0.0

## System 0.66 0.71 0.64



Physical Memory 75.2%

11.56 GB

15.38 GB

Swap Space 0.2%

15.70 GB

File Descriptor Count 0.3%

177

65536

## JVM-Memory 27.6%

135.32 MB

490.69 MB

490.69 MB



- Dashboard
- Logging
- Cloud
- Collections
- Java Properties
- Thread Dump

films

- Overview
- Analysis
- Dataimport
- Documents
- Files
- Query
- Schema

Core Selector

Request-Handler (qt)

/select

common

q

genre:Fantasy

fq

sort

start, rows

0 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☒ indent

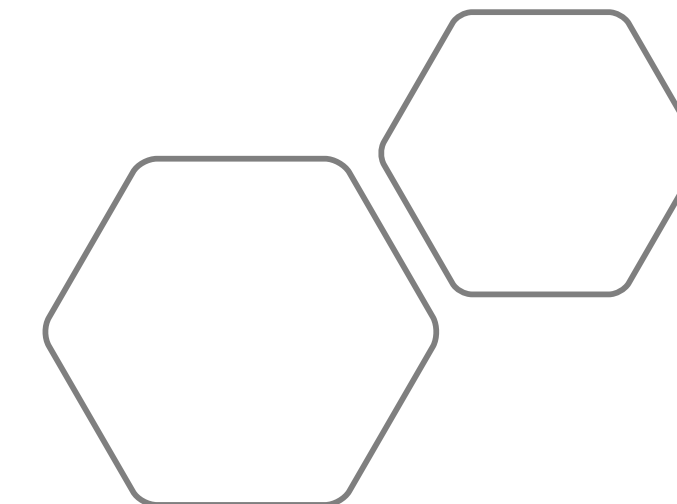
☐ debugQuery

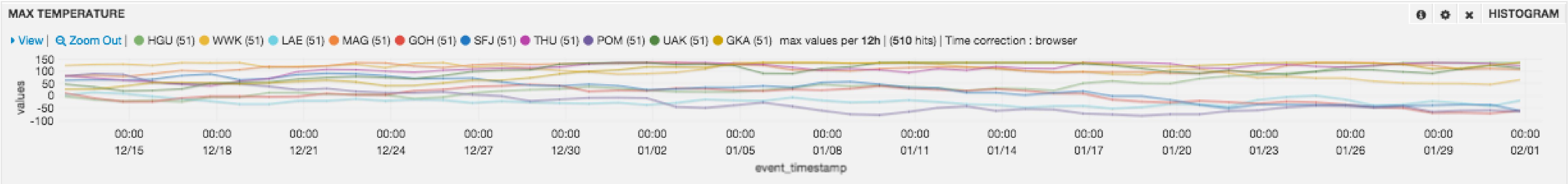
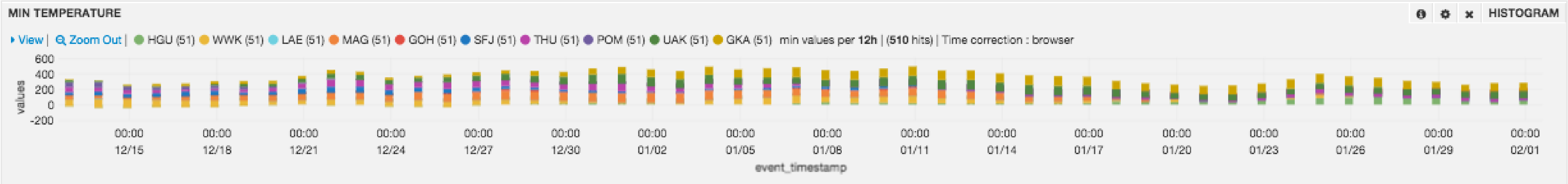
☐ dismax

☐ edismax

http://127.0.1.1:7574/solr/films/select?indent=on&q=genre:Fanta

```
{
  "responseHeader": {
    "zkConnected": true,
    "status": 0,
    "QTime": 3,
    "params": {
      "q": "genre:Fantasy",
      "indent": "on",
      "wt": "json",
      "_": "1460656348479"
    }
  },
  "response": { "numFound": 82, "start": 0, "docs": [
    {
      "id": "/en/9_2005",
      "directed_by": ["Shane Acker"],
      "initial_release_date": "2005-04-21T00:00:00Z",
      "genre": ["Computer Animation",
        "Animation",
        "Apocalyptic and post-apocalyptic fiction",
        "Science Fiction",
        "Short Film",
        "Thriller",
        "Fantasy"],
      "name": ["9"],
      "_version_": 1531518174029152256
    },
    {
      "id": "/en/300_2007",
      "directed_by": ["Zack Snyder"],
      "initial_release_date": "2006-12-09T00:00:00Z",
      "genre": ["Epic film",
        "Adventure Film",
        "Fantasy",
```





METRIC PER STATION

Limit Your Search

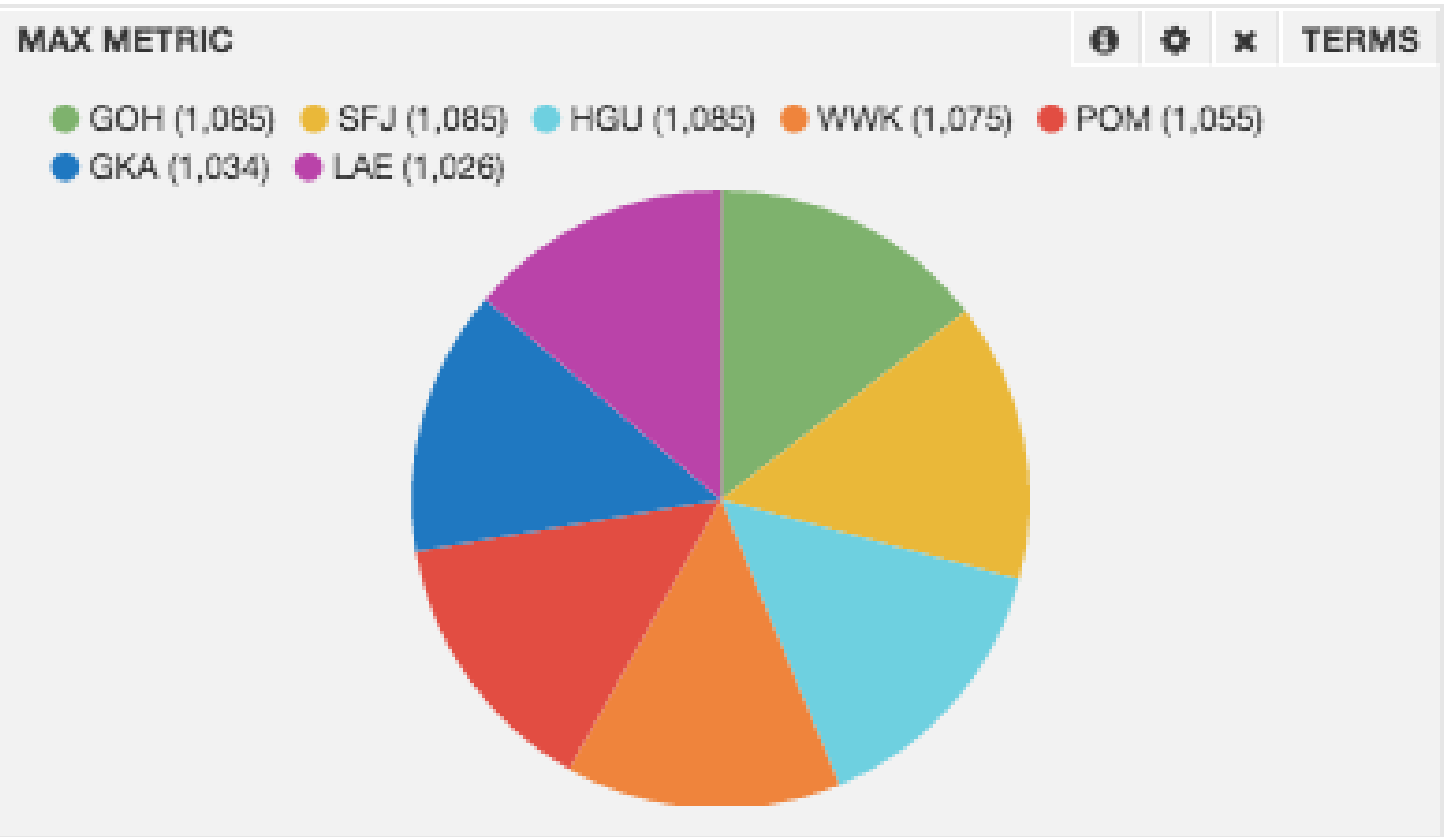
metric

- barometricpressure (510)
- dewpoint (510)
- humidity (510)
- precipitation (510)
- temperature (510)
- winddirection (510)
- windspeed (510)

METRIC RANGES

0 to 10 of 50 available for paging

event_timestamp	stationid	min	mean	max
2016-02-01T00:00:00Z	WWK	28	35	43
2016-02-01T00:00:00Z	GOH	315	327	336
2016-02-01T00:00:00Z	HGU	0	55	108
2016-02-01T00:00:00Z	POM	288	298	314
2016-02-01T00:00:00Z	GKA	999	1002	1005
2016-02-01T00:00:00Z	HGU	1062	1067	1073
2016-02-01T00:00:00Z	GOH	32	38	45
2016-02-01T00:00:00Z	SFJ	28	32	38
2016-02-01T00:00:00Z	SFJ	0	0	3



# **APACHE SOLR: PROS & CONS**

## **Điểm mạnh**

- ✓ Real-time indexing
- ✓ Rich set of features (Highlighting, “Did you mean?”...)
- ✓ Rich Content Document Support
- ✓ Data Visualization (Banana)
- ✓ Machine Learning Support
- ✓ Storage Support

## **Điểm yếu**

- × Chậm hơn Elasticsearch
- × Chỉ phù hợp với dữ liệu ít thay đổi





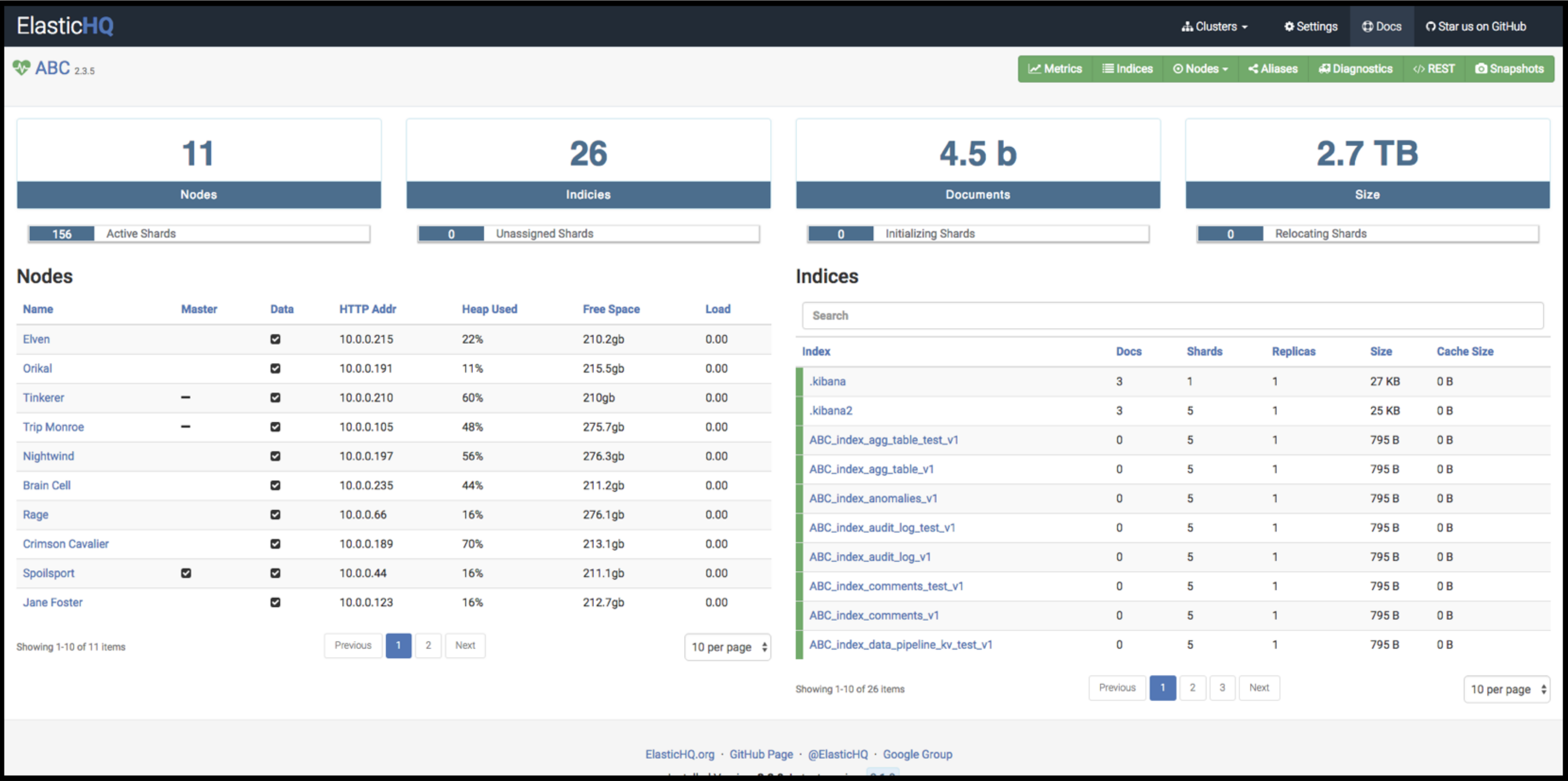
# ELASTICSEARCH: INTRODUCTION

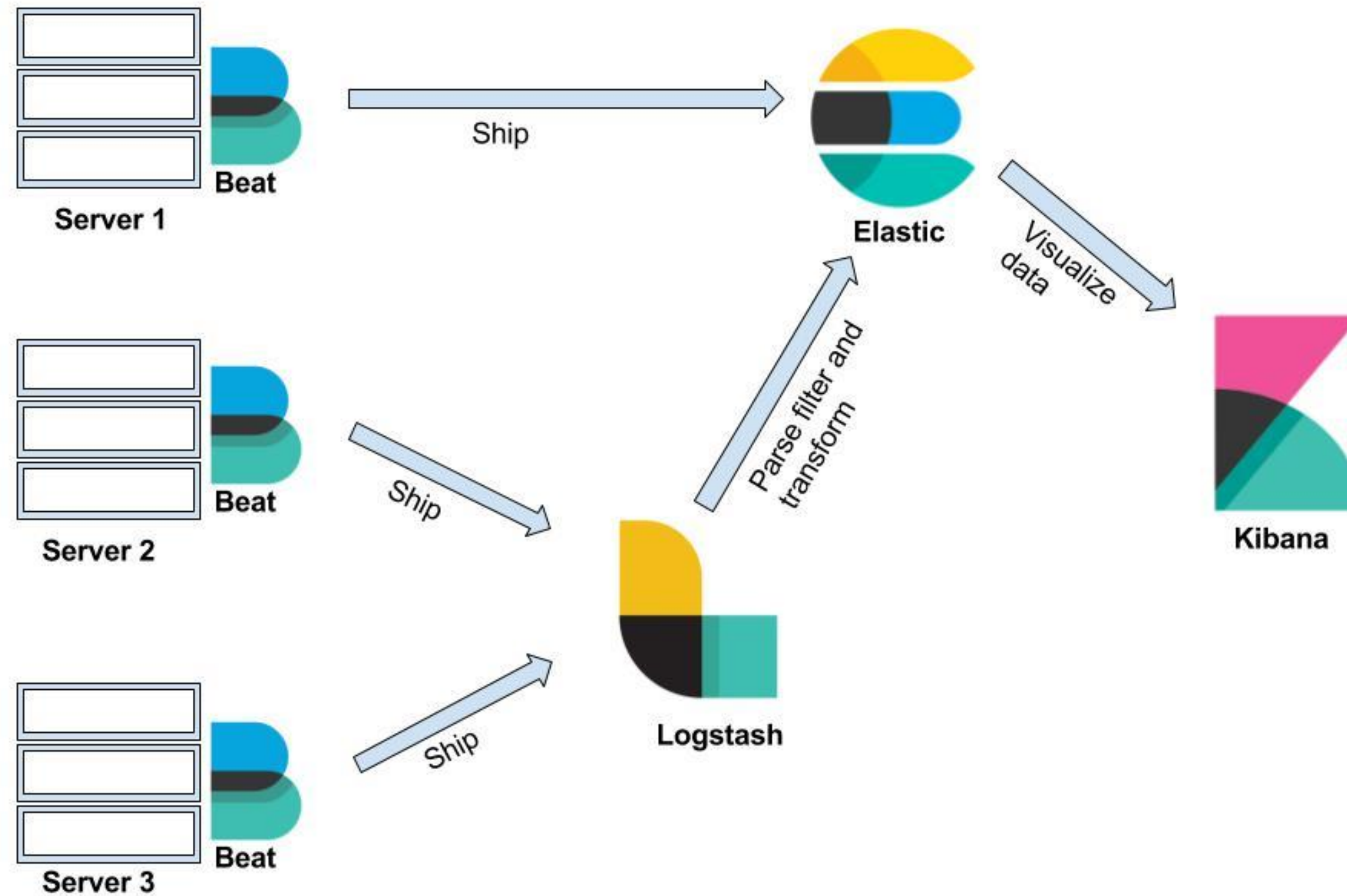
- ❑ **Elasticsearch** là một công cụ tìm kiếm và phân tích được phát triển bởi **Elastic**, phát hành phiên bản đầu tiên vào năm **2010**.
- ❑ Được phát triển bằng ngôn ngữ **Java**, dựa trên nền tảng thư viện **Apache Lucene**.
- ❑ Được sử dụng bởi nhiều công ty, tập đoàn như Cisco, eBay, Netflix, TripAdvisor, NASA, Wikipedia...
- ❑ Website: <https://www.elastic.co/elasticsearch/>



elasticsearch



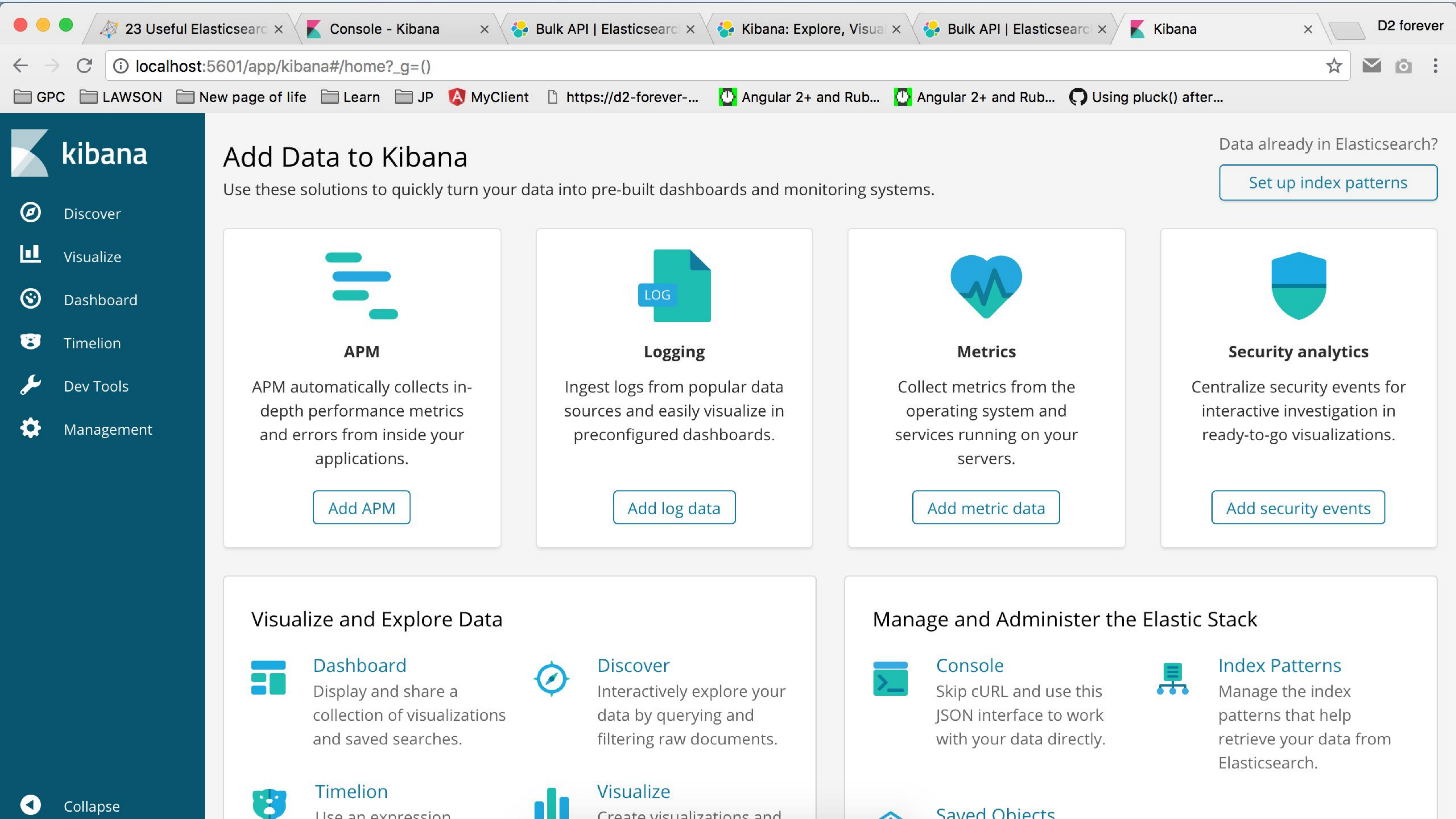




## Log and System Metrics Management with Elastic Stack







Discover



Visualize



Dashboard



Timelion



Dev Tools



Management



Collapse

## Add Data to Kibana

Use these solutions to quickly turn your data into pre-built dashboards and monitoring systems.

Data already in Elasticsearch?

[Set up index patterns](#)



### APM

APM automatically collects in-depth performance metrics and errors from inside your applications.

[Add APM](#)



### Logging

Ingest logs from popular data sources and easily visualize in preconfigured dashboards.

[Add log data](#)



### Metrics

Collect metrics from the operating system and services running on your servers.

[Add metric data](#)



### Security analytics

Centralize security events for interactive investigation in ready-to-go visualizations.

[Add security events](#)

## Visualize and Explore Data



### Dashboard

Display and share a collection of visualizations and saved searches.



### Discover

Interactively explore your data by querying and filtering raw documents.



### Timelion

Use an expression



### Visualize

Create visualizations and

## Manage and Administer the Elastic Stack



### Console

Skip cURL and use this JSON interface to work with your data directly.



### Index Patterns

Manage the index patterns that help retrieve your data from Elasticsearch.



### Saved Objects



## Console

```
1 GET /bookdb_index/book/_search
```

```
2
3 {
4     "query": {
5         "match" : {
6             "query" : "guide",
7             "fields" : ["_all"]
8         }
9     }
10 }
```

```
1 {
2   "took": 1,
3   "timed_out": false,
4   "_shards": {
5     "total": 1,
6     "successful": 1,
7     "skipped": 0,
8     "failed": 0
9   },
10  "hits": {
11    "total": 4,
12    "max_score": 1,
13    "hits": [
14      {
15        "_index": "bookdb_index",
16        "_type": "book",
17        "_id": "1",
18        "_score": 1,
19        "_source": {
20          "title": "Elasticsearch: The Definitive Guide",
21          "authors": [
22            "clinton gormley",
23            "zachary tong"
24          ],
25          "summary": "A distributed real-time search and analytics engine",
26          "publish_date": "2015-02-07",
27          "num_reviews": 20,
28          "publisher": "oreilly"
29        }
30      },
31      {
32        "_index": "bookdb_index",
33        "_type": "book",
```

# **ELASTICSEARCH: PROS & CONS**

## **Điểm mạnh**

- ✓ Real-time Indexing
- ✓ High Scalability
- ✓ Storage Support
- ✓ Visualization of Data (Elastic Stack)
- ✓ Security Analytics
- ✓ Machine Learning Support
- ✓ Cloud Service Support

## **Điểm yếu**

- × Công nghệ tuổi đời còn thấp (so với Apache Solr và Sphinx) nên còn thiếu nhiều tính năng so với đối thủ





# **SPHINX: INTRODUCTION**

- ❑ **Sphinx** là một công cụ tìm kiếm mã nguồn mở hỗ trợ tìm kiếm cho nhiều Database Management System như **MySQL, PostgreSQL...**
- ❑ Sphinx được phát triển bằng ngôn ngữ **C++**, phát triển bởi **Sphinx Technology Inc.** Phát hành phiên bản đầu tiên vào năm **2001**.
- ❑ Được sử dụng bởi một số công ty, tập đoàn như Joomla, Tumblr, Envato...
- ❑ Website: <http://sphinxsearch.com/>

















# **SPHINX: INTRODUCTION**

- ❑ **Sphinx** là một công cụ tìm kiếm mã nguồn mở hỗ trợ tìm kiếm cho nhiều Database Management System như **MySQL, PostgreSQL...**
- ❑ Sphinx được phát triển bằng ngôn ngữ **C++**, phát triển bởi **Sphinx Technology Inc.** Phát hành phiên bản đầu tiên vào năm **2001**.
- ❑ Được sử dụng bởi một số công ty, tập đoàn như Joomla, Tumblr, Envato...
- ❑ Website: <http://sphinxsearch.com/>



- Node Dashboard
- Nodes
- Template Indexes
- Plain Indexes
- RealTime Indexes
- Distributed Indexes

## SphinxQL on forum

[Browse](#)[Schema](#)[Search](#)[SphinxQL](#)[Token test](#)[Excerpts](#)[Operations](#)[Stats](#)

```
SELECT *,WEIGHT() as relevance FROM forum WHERE MATCH('Ranking exact matches ') ORDER BY relevance DESC LIMIT 0,10
OPTION ranker=sph04,idf=plain
```

SELECT \*SELECTUPDATEClear

Enable ProfilingSHOW PLANSHOW META

Execute

author_id ↓	forum_id ↓	id ↓	relevance ↓	subject ↓	
4172	1	4519	22609	Ranking exact matches first	<div>UpdateView</div>
2078	1	6963	16597	extended query syntay: match exact match	<div>UpdateView</div>
6945	1	1923	14628	Exact result first	<div>UpdateView</div>
27160	1	11436	14613	Exact match for ranking	<div>UpdateView</div>
46364	1	11379	14595	Exact match ranking and weights	<div>UpdateView</div>
26702	1	9508	14580	Exact match not shown first	<div>UpdateView</div>
26048	1	8236	14576	exact match ranked higher	<div>UpdateView</div>
4775	1	4521	12615	Weight highest when exact match	<div>UpdateView</div>
4470	1	3344	12606	Basic Ranking Question & Adding Special	<div>UpdateView</div>
26045	1	8593	12605	SPH_RANK_SPH04 Ranking broken?	<div>UpdateView</div>

⏮⏪12345678⏩⏭



# **SPHINX: PROS & CONS**

## **Điểm mạnh**

- ✓ Powerful and Fast (300 triệu query / day, indexing lên đến 25 tỷ record)
- ✓ Real-time indexing
- ✓ Nothing Useless

## **Điểm yếu**

- × Không phù hợp với các dữ liệu không có cấu trúc (DOCs, PDFs, MP3s...)
- × Tốn nhiều công sức để thiết lập



# COMPARISON

	Apache Solr	Elasticsearch	Sphinx
Tính năng	<ul style="list-style-type: none"> <li>• Full-text</li> <li>• Autocomplete Suggestion</li> <li>• Faceted</li> <li>• Multifield</li> <li>• Synonyms</li> <li>• Fuzzy</li> <li>• Geospatial</li> </ul>	<ul style="list-style-type: none"> <li>• Full-text</li> <li>• Autocomplete Suggestion</li> <li>• Faceted</li> <li>• Multifield</li> <li>• Synonyms</li> <li>• Fuzzy</li> <li>• Geospatial</li> <li>• Highlighting</li> <li>• Spell Checker</li> </ul>	<ul style="list-style-type: none"> <li>• Full-text</li> <li>• Autocomplete Suggestion</li> <li>• Faceted</li> <li>• Multifield</li> <li>• Synonyms</li> <li>• Geospatial</li> <li>• Highlighting</li> <li>• Spell Checker</li> </ul>
Real time indexing	Có	Có	Có
Hiệu suất	Cao	Cao	Cao
Khả năng mở rộng	Cao	Cao	Cao



# COMPARISON

	Apache Solr	Elasticsearch	Sphinx
Visualization	Elastic Stack	Banana	Không hỗ trợ
Machine Learning	Có	Có	Không hỗ trợ
Storage	Có	Có	Không hỗ trợ
SaaS	Không hỗ trợ	Có	Không hỗ trợ
Java API RESTful API	Có	Có	Không hỗ trợ
Transaction	Optimistic Locking	Không hỗ trợ	Không hỗ trợ
Programming Language	<ul style="list-style-type: none"> <li>• .NET</li> <li>• Java</li> <li>• JavaScript</li> <li>• PHP</li> <li>• Python</li> <li>• Ruby</li> <li>• Perl</li> </ul>	<ul style="list-style-type: none"> <li>• .NET</li> <li>• Java</li> <li>• JavaScript</li> <li>• PHP</li> <li>• Python</li> <li>• Ruby</li> <li>• Perl</li> </ul>	<ul style="list-style-type: none"> <li>• C++</li> <li>• Java</li> <li>• PHP</li> <li>• Python</li> <li>• Ruby</li> <li>• Perl</li> </ul>





# APPLYING

Apache Solr	Elasticsearch	Sphinx
<ul style="list-style-type: none"><li>• Dữ liệu ít thay đổi</li><li>• Dữ liệu dạng tài liệu, hình ảnh...</li><li>• Có nhu cầu tích hợp ML/AI</li><li>• Có nhu cầu visualize dữ liệu</li></ul>	<ul style="list-style-type: none"><li>• Hầu hết mọi nhu cầu</li><li>• Có nhu cầu tích hợp ML/AI</li><li>• Có nhu cầu visualize dữ liệu</li><li>• Có nhu cầu sử dụng Cloud</li></ul>	<ul style="list-style-type: none"><li>• Dữ liệu rất lớn</li><li>• Mục đích duy nhất là tìm kiếm dữ liệu</li></ul>



ありがとうございます

