

# Mining Modern Repositories with Elasticsearch

Zhiyuan Lin

# ElasticSearch



- created by Shay Banon
- released in 2010
- The company elastic was founded in 2012 to provide commercial solution around ES and related software.

# ElasticSearch



ElasticSearch is a great open source search engine built on top of Apache Lucene.

The Apache Lucene™ project develops open-source search software.

The screenshot shows the Elastic website's search results page. At the top, the Elastic logo and name are displayed. A navigation menu includes links for PRODUCTS, FOUND, SUBSCRIPTIONS, LEARN, COMMUNITY, USE CASES, BLOG, and ABOUT. On the right, there are links for downloads, docs, support, and discuss, along with a search bar containing the text "elastic(search)".

The main content area features a search result titled "Found: Hosted Elasticsearch". Below the title, it says "SPIN UP A CLUSTER IN SECONDS" and "The only hosted Elasticsearch service supported and managed by Elastic." There are two buttons: "Learn More" and "Free Trial". To the right of the text is a cartoon illustration of a blue dog wearing a red collar with a colorful tag, holding a magnifying glass.

At the bottom of the page, there are logos for LinkedIn, Wikipedia (The Free Encyclopedia), TaskRabbit, WordPress, and FICO.

# Back to the paper...

## Main Idea?

- not a typical paper
- doesn't come up with any new ideas
- simply evaluates Elasticsearch based on a tool called "Dash."

# Why do we need ES?

- Companies are generating data that often exceeds their ability to analyze
- Insights derived from large data sets are crucial
- Impractical to analyze using traditional databases.

➔ Elasticsearch

# ElasticSearch

(It is an open source full-text search engine written in Java that is designed to be :)

- Distributive
  - Automatic sharding
  - Automatic distribution of shards among the nodes in a cluster
- Scalable
- Near real-time capable
  - Each shard is being indexed/refreshed independently
  - indices are constantly refreshed with fixed time interval

# Elasticsearch v.s. RDBMS

- All data in ES is stored in “indices”
- Each document in ES is a JSON object, analogous to a row in a table in a RDBMS.
- Document type defines the set of fields that can be specified for a particular document.

**Table 1: Elasticsearch vs. SQL**

Elasticsearch element	SQL element
Index	Database
Mapping	Schema
Document type	Table
Document	Row

# The distributed nature of ES

- Automatic sharding

(Each ES index consists of one or more Lucene indices, called shards. )

(automatically defines the shard that will be responsible for storing and indexing the new document )

- Automatic distribution of shards among the nodes in a cluster

Eg.-An index consists of six shards

- the cluster only has one node

- all shards are on the same node

- add one more node to the cluster

- automatically move half of the shards to the new node



# Elasticsearch

- **Communication with Server**

(as long as the client can send HTTP requests)

- **Mapping**

(similar to the schema definition in SQL databases.  
defines all document types within the index.)

- **Near Real-time search**

(ES server does not refresh indices after each update,  
instead, it uses a specified fixed time interval to refresh.)

# Performing a search

Elasticsearch provides its own query language based on JSON called Query DSL.

To execute a query, a client sends a search request to (one of the following) addresses:

```
http://<server>/_search
```

```
http://<server>/<index>/_search
```

```
http://<server>/<index>/<documentType>/_search
```

```
{ "query":{ "filtered":{  
  "query":{ "match_all":{ } },  
  "filter":{ "and":[  
    { "range":  
      { "modified_ts":  
        { "gte":0, "lt":1400000000000 } } },  
    { "term":  
      { "reported_by": "johndoe@mozilla.com" } } },  
    { "terms":{  
      "bug_status": [ "new", "reopened" ] } } },  
    { "not":{ "term":{ "priority": "p1" } } }  
  ] }  
} },  
"from":0,  
"size":100,  
"fields":["bug_id"] }
```

**Figure 1: A sample search query using filters.**

# Evaluation

- Software analytics

(A developer dashboard tool--DASH. )

- first implemented using Bugzilla's REST API

- big improvement in the execution time, and the average response time of querying(after switching to ES).

- Social Media Analysis...

(Facebook, Github, Netflix, Stack Overflow.....)

# Weaknesses

<https://www.quora.com/Why-shouldnt-I-use-Elasticsearch-as-my-primary-datastore>

- **Learning Curve**

- easy to start writing simple queries
- query writing becomes more complicated  
if it involves nested object.

- **Security**

- Elasticsearch does not provide any authentication  
or access control functionality.

(If someone knows the url of the server, he can easily delete all the indices and shut down the server)

(when searching “impact of es” in BAIDU)

-(weaknesses of being NoSQL system - lack transactions, lack of JOIN operation, possible inconsistencies in data, etc.)

# Related Works

- From ES's Wikipedia:

Elasticsearch is the second most popular enterprise search engine after Apache Solr.

- Solr

Solr is a standalone enterprise search server with a REST-like API.

-Both of them are in the Lucene family.

-Which one to use???

here is the answer!!!

If you are working on a small app that needs to search less than a million documents and the database doesn't update much. Solr is the choice.

# Discussion

- The future of elasticsearch?
- Shay is active,
- The project is active(adding features).
- **Other blogs** -I found that ES has certainly already come a long way since its creation in 2010, and now offers far more than simple search. Not only has the scope of the project expanded but integration has improved.
- **In the past five years**
  - the software has been downloaded around 20 million times
  - the company has raised over \$100m;