**Elasticsearch, Kibana, Logstash**

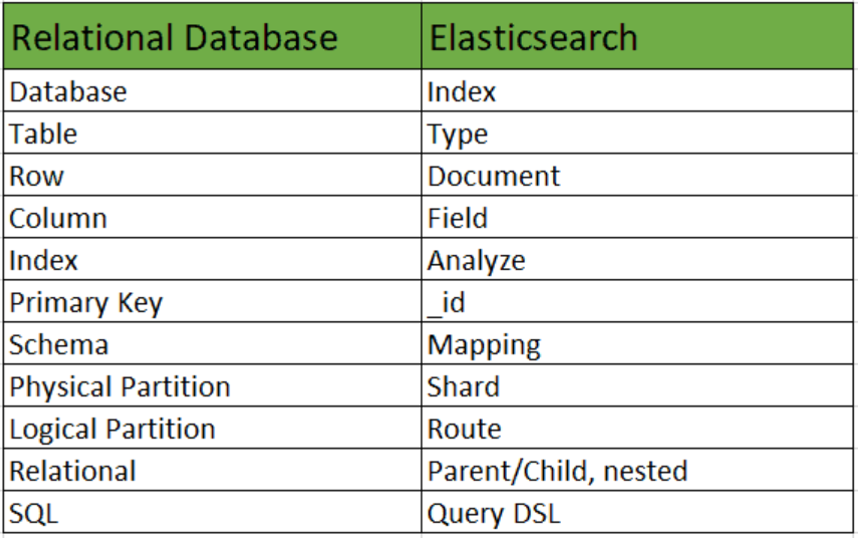
1. Tổng quan
2. Elasticsearch

Khái niệm:

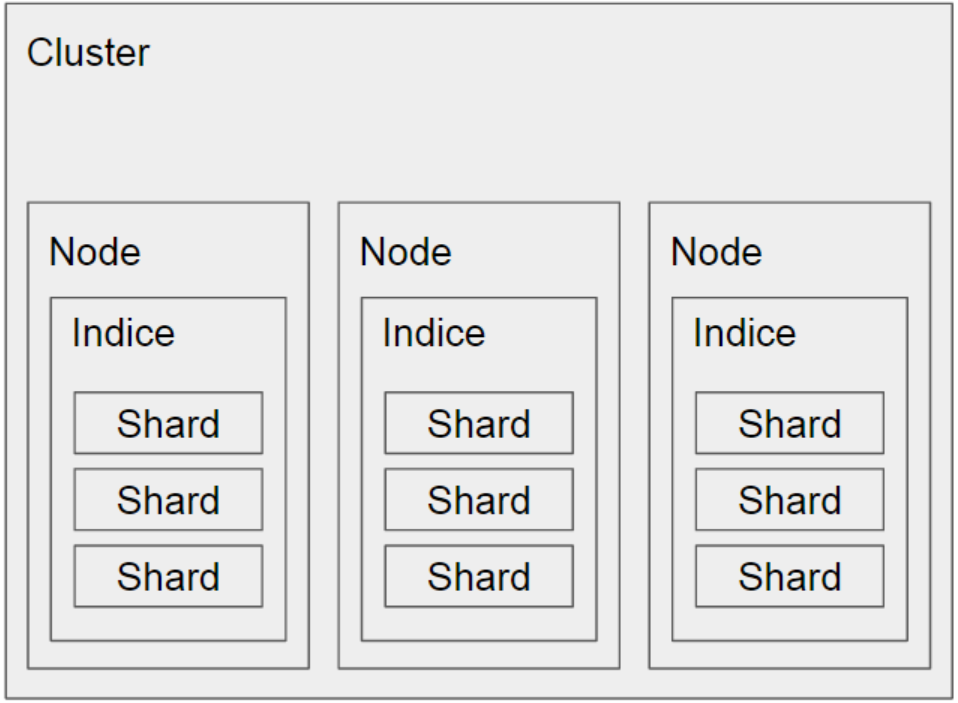
* Elasticsearch là một công cụ tìm kiếm dựa trên nền tảng Apache Lucene. Nó cung cấp một bộ máy tìm kiếm dạng phân tán, có đầy đủ công cụ với một giao diện web HTTP có hỗ trợ dữ liệu JSON.
* Elasticsearch là một search engine được kế thừa từ Lucene Apache động như một web server, có khả năng tìm kiếm nhanh chóng (near realtime) thông qua giao thức RESTful, có khả năng phân tích và thống kê dữ liệu.
* Elasticsearch là 1 hệ thống phân tán và có khả năng mở rộng tuyệt vời (horizontal scalability). Lắp thêm node cho nó là nó tự động mở rộng kích thước dữ liệu.
* Elasticsearch là 1 open source được phát triển bằng Java.

1. Các khái niệm cơ bản trong Elasticsearch

* Cluster: Elasticsearch hoạt động trong một cụm gọi là cluster. Một cluster Elasticsearch bao gồm một hoặc nhiều node, là các máy chủ Elasticsearch đang chạy. Cluster giúp phân tải công việc và cung cấp khả năng chịu lỗi.
* Node: Một node là một instance của Elasticsearch chạy trên một máy chủ. Mỗi node trong cluster có vai trò đóng vai trò cụ thể trong việc xử lý và lưu trữ dữ liệu.
* Index: Một index là một tập hợp các tài liệu có cùng cấu trúc được nhóm lại trong Elasticsearch để tìm kiếm và truy xuất dữ liệu. Index được chia thành các shard (đoạn) nhằm phân tải và phân chia dữ liệu trên các node khác nhau trong cluster.
* Shard: Một shard là một phần dữ liệu của một index. Elasticsearch chia index thành nhiều shard để phân tải và phân chia công việc xử lý trên nhiều node. Có hai loại shard: primary shard (đoạn chính) và replica shard (đoạn sao chép). Mỗi primary shard có một bản sao replica shard để đảm bảo sự dự phòng và khả năng chịu lỗi.
* Document: Một document là một bản ghi dữ liệu JSON trong Elasticsearch. Mỗi document thuộc vào một index và có một ID duy nhất. Elasticsearch sử dụng cấu trúc dữ liệu JSON để lưu trữ và truy vấn document.
* Query: Elasticsearch cung cấp một ngôn ngữ truy vấn linh hoạt để tìm kiếm và truy xuất dữ liệu. Người dùng có thể sử dụng các truy vấn để chỉ định tiêu chí tìm kiếm, lọc và sắp xếp dữ liệu.
* MAPPING là quá trình xử lý cách mà các DOCUMENT (và các PROPERTIES bên trong) sẽ được index và lưu trữ như thế nào. MAPPING giúp chúng ta cùng lúc khởi tạo 1 field & định nghĩa cách field các thuộc tính của trường trong index.



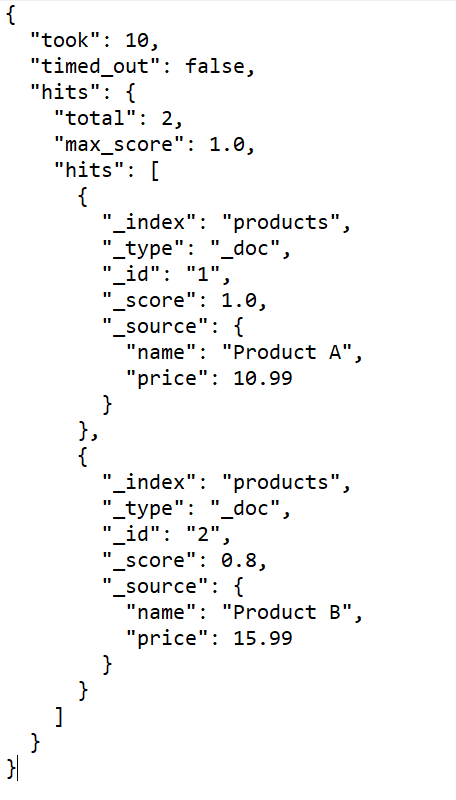
Ảnh 1.0 so sánh Es so với CSDL



Ảnh 1.1 Cấu trúc của ES

1. Dữ liệu trong ES

* Trong Elasticsearch dữ liệu được lưu trữ thành các tài liệu dưới dạng JSON (key, value).
* Cấu trúc dạng dữ liệu của Elasticsearch được xác định theo cơ mapping cho các trường dữ liệu.

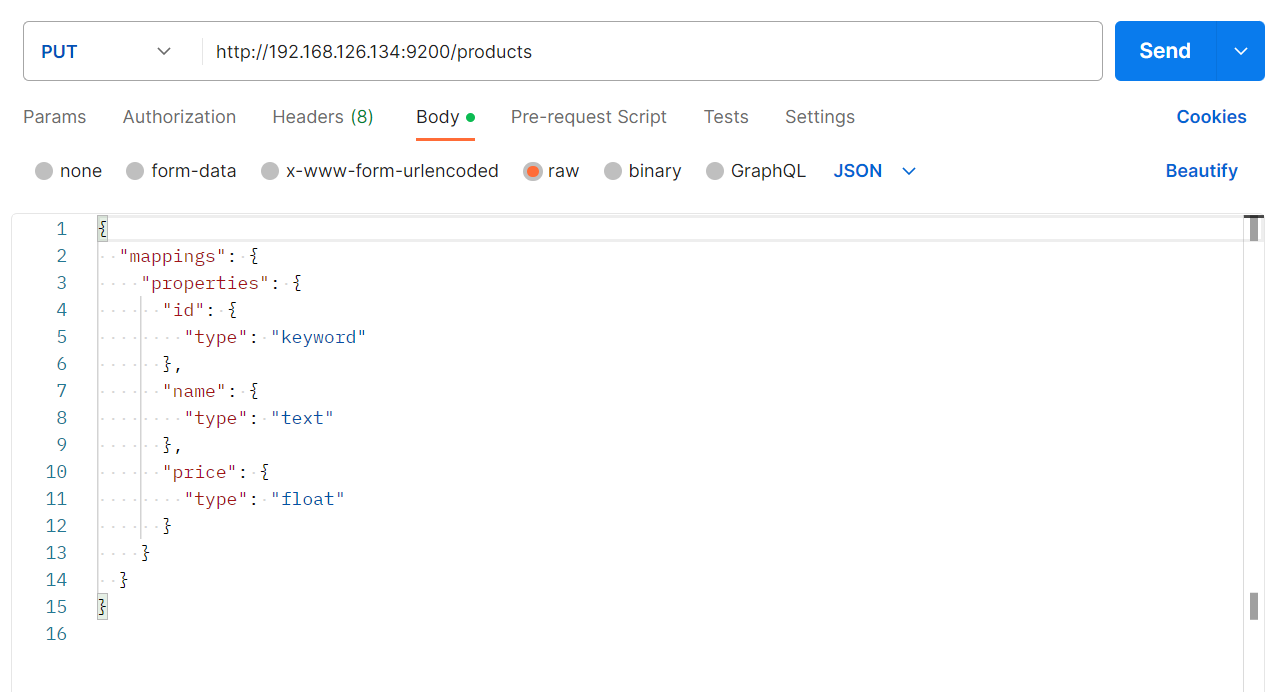
Ảnh 1.2 Mapping và dữ liệu trong ES.

1. Ưu nhược điểm của ES

* Ưu điểm:
  1. Tìm kiếm mạnh mẽ: ES cung cấp khả năng tìm kiếm toàn văn bản nhanh chóng và chính xác. Nó sử dụng cơ chế inverted index và tf-idf (term frequency-inverse document frequency) để cải thiện hiệu suất tìm kiếm.
  2. Phân tán và mở rộng: ES được thiết kế để hoạt động trong môi trường phân tán, cho phép phân tải công việc và dữ liệu trên nhiều node. Điều này giúp ES có khả năng mở rộng tốt, xử lý khối lượng lớn dữ liệu và đáp ứng tải cao.
  3. Khả năng xử lý dữ liệu phong phú: ES cung cấp các tính năng phân tích dữ liệu phong phú như tổng hợp, phân loại, gợi ý và khám phá dữ liệu. Nó tích hợp các công cụ phân tích mạnh mẽ như Kibana để hỗ trợ việc trực quan hóa và phân tích dữ liệu.
  4. Tích hợp linh hoạt: ES tích hợp tốt với hệ sinh thái phân tích và giao diện RESTful cho phép tương tác dễ dàng với dữ liệu. Nó có khả năng tích hợp với các công cụ và framework phổ biến như Logstash, Beats và Kafka.
  5. Độ tin cậy và khả năng phục hồi: ES sử dụng cơ chế phân tán và sao chép để đảm bảo khả năng chịu lỗi và độ tin cậy cao. Dữ liệu được sao chép trên nhiều node, cho phép khôi phục dữ liệu trong trường hợp sự cố.
* Nhược điểm
  1. Tính phức tạp của hệ thống: Elasticsearch là một hệ thống phân tán mạnh mẽ, nhưng cũng có sự phức tạp trong việc triển khai và quản lý. Yêu cầu kiến thức và kỹ năng để xử lý và điều chỉnh hệ thống một cách hiệu quả.
  2. Tính năng SQL hạn chế: Mặc dù ES cung cấp một số tính năng SQL thông qua plugin, nhưng nó không hỗ trợ đầy đủ ngôn ngữ truy vấn SQL. Điều này có thể gây khó khăn cho những người đã quen thuộc với SQL trong việc tương tác với ES.
  3. Tiêu thụ tài nguyên cao: ES yêu cầu tài nguyên phần cứng và bộ nhớ để hoạt động hiệu quả. Các cluster ES có thể tiêu thụ nhiều tài nguyên hơn so với một hệ thống cơ sở dữ liệu truyền thống, đặc biệt là khi xử lý các tác vụ tìm kiếm và phân tích phức tạp.
  4. Khả năng phân tích phụ thuộc vào cấu trúc dữ liệu: ES là một cơ sở dữ liệu không cấu trúc, điều này có nghĩa là việc phân tích dữ liệu phụ thuộc vào cấu trúc và mapping đúng. Nếu mapping không được định nghĩa chính xác, việc tìm kiếm và phân tích có thể không hiệu quả.

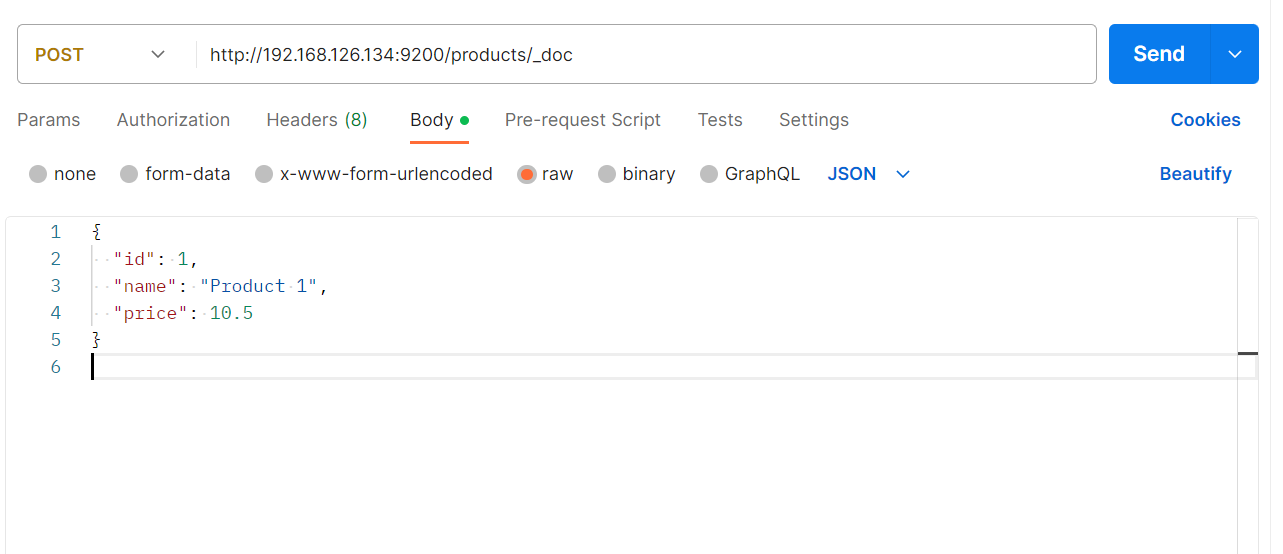
1. Các lệnh cơ bản trong ES

* Tạo index:
  1. Sử dụng câu lệnh với HTTP PUT hoặc POST. Thao thác trên cmd sử dung curl hoặc thao tác trên postman



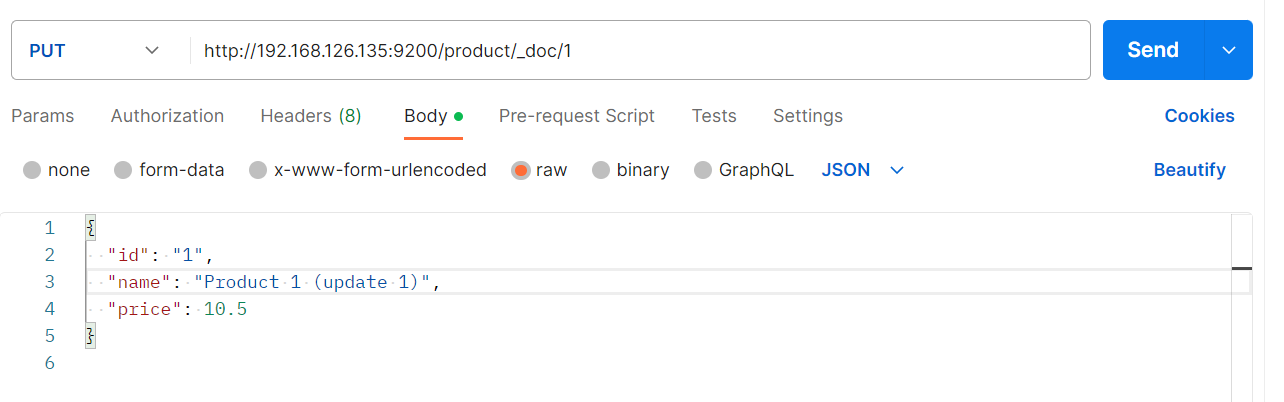
Ảnh 1.3 Thao tác tạo index với postman.

* Thêm dữ liệu vào trong index
  1. Sử dụng put hoặc post truyền json data sao cho dữ liệu data mapping với cấu trúc mapping của index (lưu ý về url của việc insert). Url của ví dụ: <http://192.168.126.134:9200/products/_doc>
  2. Có thể insert một mảng data vào trong Elasticsearch.



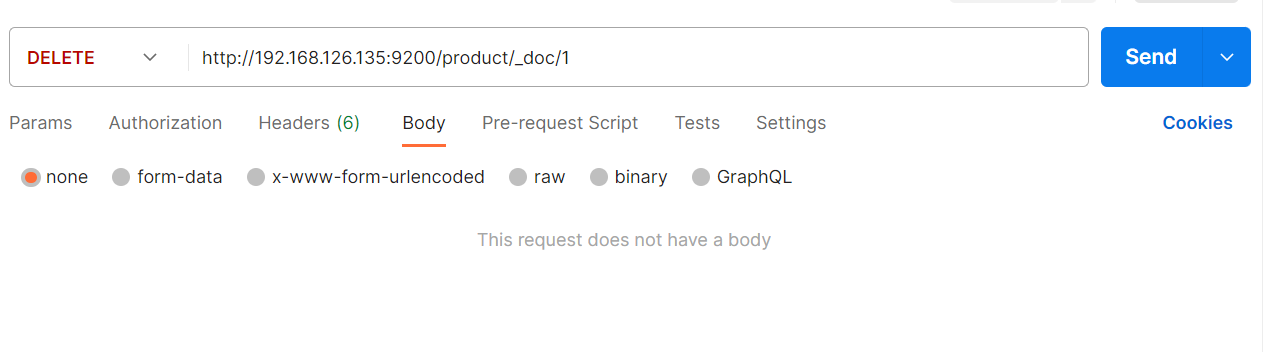
Ảnh 1.4 Thao tác insert data.

* Chỉnh sửa dữ liệu
  1. Để chỉnh sửa dữ liệu ta dùng PUT để thao tác chỉnh sửa với cấu trúc URI: **[Error! Hyperlink reference not valid.](http://<IP)** Elasticsearch>:port/<index name>/\_doc/id



Ảnh 1.5 Chỉnh sửa dữ liệu

* Xoá dữ liệu
  1. Thao tác xoá dữ liệu với DELETE method với uri **[Error! Hyperlink reference not valid.](http://<IP)** Elasticsearch>:port/<index name>/\_doc/id

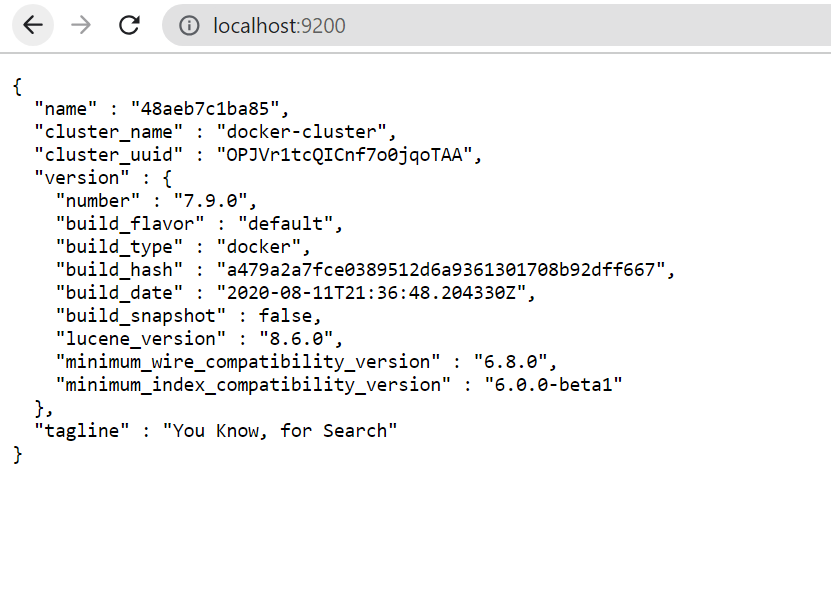


Ảnh 1.6 Xoá dữ liệu

* Thao tác tìm kiếm trong ES
  1. Elasticsearch cung cấp nhiều loại query khác nhau để phù hợp với các nhu cầu tìm kiếm khác nhau như
     + Match Query: Đây là query phổ biến nhất trong Elasticsearch. Nó tìm kiếm các documents có trường chỉ định chứa một giá trị cụ thể hoặc một từ khóa. Query này sử dụng phương pháp tìm kiếm full-text để tìm kiếm và tính điểm độ trùng khớp của các documents.
     + Term Query: Query này tìm kiếm các documents có trường chỉ định chứa một giá trị cụ thể mà không thực hiện tính điểm. Nó dùng chính xác từ khóa để tìm kiếm.
     + Range Query: Query này tìm kiếm các documents có trường chỉ định nằm trong một khoảng giá trị cụ thể. Khoảng có thể là khoảng giá trị số, ngày tháng, hoặc thậm chí là khoảng giá trị của một trường text được sắp xếp theo thứ tự lexicographic.
     + Bool Query: Query này kết hợp nhiều query khác nhau bằng các toán tử logic (AND, OR, NOT) để tạo ra các tập hợp điều kiện tìm kiếm phức tạp.
     + Nested Query: Query này cho phép tìm kiếm các documents dựa trên các cặp trường-giá trị nằm trong một trường dữ liệu được nhúng (nested) trong document.
     + Elasticsearch còn cung cấp nhiều loại query khác nhau như Wildcard Query, Fuzzy Query, Phrase Query, và nhiều loại query khác để phục vụ các nhu cầu tìm kiếm đa dạng.
     + Tìm hiểu thêm các câu query truy cập: <https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl.html>

1. Cài đặt Elasticsearch

* Cài đặt Elasticsearch với docker
  1. Sử dụng docker để pull image elasticsearch với các version cần sử dụng.Trong ví dụ tôi sử dụng Elasticsearch version 7.9.0 tham khảo trên https://hub.docker.com/\_/elasticsearch
  2. Câu lệnh: **docker network create somenetwork**
  3. Sau khi pull thành công chạy lệnh: **docker run -d --name elasticsearch --net somenetwork -p 9200:9200 -p 9300:9300 -e "discovery.type=single-node" elasticsearch:7.9.0**
  4. Sau đó truy cập địa chỉ <http://localhost:9200> hiện thông tin elasticsearch



Ảnh 1.7 Truy cập elasticsearch

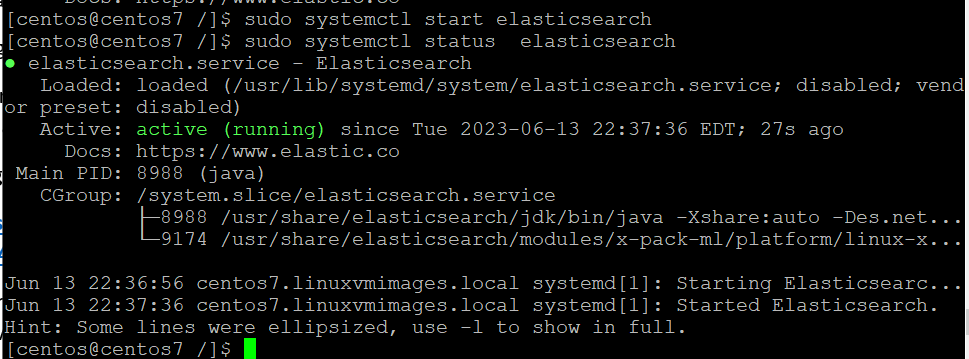
* 1. Sau khi truy cập thành công ta có thể dùng post man để thao tác với elasticsearch.
  2. Việc cấu hình elasticsearch trên docker sẽ tuỳ vào việc cấu hình trên file **elasticsearch.yml** của bạn. Khi lệnh run ta dùng thêm option -v <vị trí file elastic>:<vị trí trên máy ảo>. Ví dụ: docker run -d -p 9200:9200 -p 9300:9300 -v D:/abc/elasticsearch.yml:/usr/share/elasticsearch/config/elasticsearch.yml elasticsearch:7.9.0 trong ví dụ này file config sẽ nằm ở ổ D và thay cho config của ES
* Cài đặt trên windown.
  1. Tuỳ từng phiên bản Elasticsearch sẽ cần các bản jdk java khác nhau. Đối với phiên bản Elasticsearch 7.9.0 khuyến nghị nên cài java 11 [Java Archive Downloads - Java SE 11 (oracle.com)](https://www.oracle.com/java/technologies/javase/jdk11-archive-downloads.html)
  2. Tiến hành tạo JAVA\_HOME trên Path environment của Windown.
  3. Download file rar của Elasticsearch trong ví dụ dùng 7.9.0 theo link <https://www.elastic.co/fr/downloads/past-releases/elasticsearch-7-9-0>
  4. TIến hành giải nén và mở thư mục path\elasticsearch-7.9.0\bin
  5. Sau cùng chạy file **elasticsearch.bat** để khởi chạy ES. Tuỳ từng phiên bản sẽ yêu cầu các jdk có version khác nhau. File config được lưu \elasticsearch-7.9.0\config
* Cài đặt trên máy ảo linux centos 7

Lưu ý có rất nhiều cách cài đặt Elasicsearch trên linux dưới đây là cách cài đặt đối với centos Elasticsearch 7.9.0.

* 1. Sau khi máy linux được khởi động
     + tiến hành update hệ điều hành với cầu lệnh: “**sudo yum update ”.**
     + Sau đó tiến hành cài java 11 cho máy “ **yum install java-11-openjdk-devel** “ hoặc “ **sudo yum install -y java-11-openjdk-devel**” để bỏ qua bước xác thực cài đặt.
     + Kiểm tra phiên bản java với “ **java -version** “.
  2. Cài đặt Elasticsearch có rất nhiều cách như cài đặt qua repo, cài đặt wget để tải xuống Elasticsearch.
     + Thêm gói wget trên centos “ **sudo yum install wget** ”
     + Lưu ý về vị trí lưu file wget phải đủ dung lượng tránh việc không đủ không gian để tải dẫn đến thiếu file khi cài đặt
     + Sau đó dùng lệnh

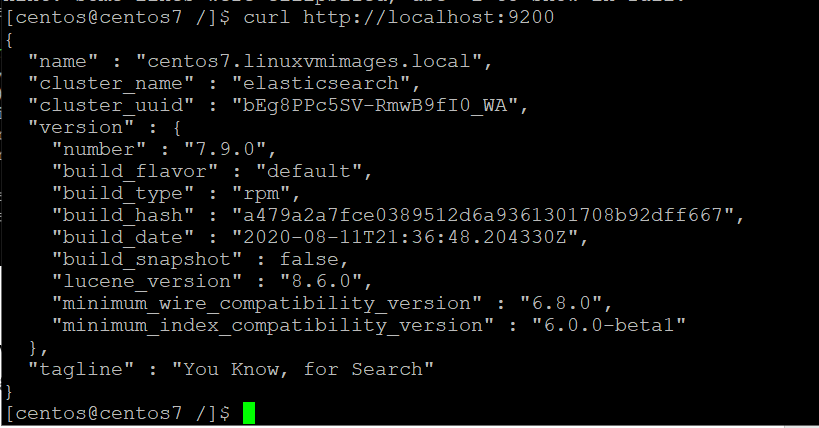
“ **wget** [**https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.9.0-x86\_64.rpm**](https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-7.9.0-x86_64.rpm)” để tải xuống gói 7.9.0

* + - Dùng câu lệnh “ **sudo rpm -ivh elasticsearch-7.9.0-x86\_64.rpm “** để tiến hành cài đặt gói. Với rpm là công cụ quản lý gói trên CentOS.
    - Sau khi cài đặt xong tiến hành kiểm tra với “ **sudo systemctl status elasticsearch** ”. Tiến hành khởi động kiểm tra với lệnh “ **sudo systemctl start elasticsearch** ” hoặc “ **sudo service elasticsearch start “**



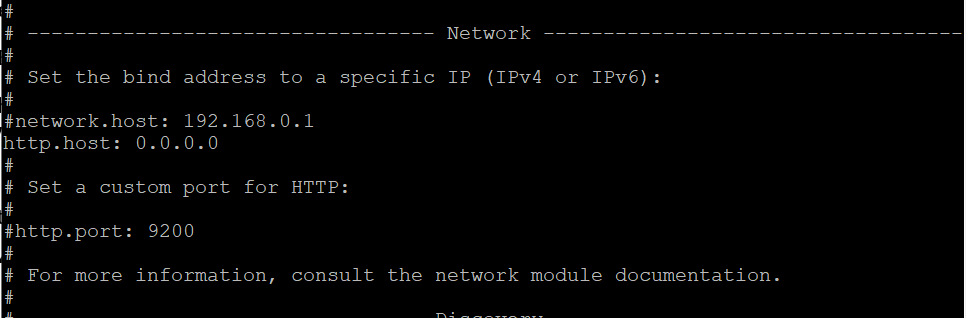
Ảnh 1.8 Chạy và kiểm tra trạng thái ES.

* + - Để kiểm tra về hoạt động của ES dùng câu lệnh “ **curl** [**http://localhost:9200**](http://localhost:9200) “ hiện ra thông tin của Elasticsearch.



Ảnh 1.9 Thông tin về Elasticsearch.

* 1. Tiến hành config cho elasticsearch
     + Cấu hình được lưu ở file elasticsearch.yml nằm trong đường dẫn “ **/etc/elasticsearch/elasticsearch.yml** ”. Để xem thông tin cài đặt ta có thể dùng lệnh **cat** với câu lệnh “ **sudo cat /etc/elasticsearch/elasticsearch.yml** “.
     + Để tiến hành chỉnh sửa file dùng VIM hoặc NANO. Trong phần cài đặt ta cần chú ý một vài điểm cơ bản như **node.name, cluster.name, http.port, path.data, path.log …**
     + Để chỉnh sửa về phầm máy ảo ta chỉnh sửa trong file **jvm.options** với đường đẫn “ **/etc/elasticsearch/jvm.options** “. Cần lưu ý một vài thông tin như setup kích thước hiệu suất **Xms, Xmx** là kích thước giới hạn bộ nhớ mặc định là 1g ta có thể thay thế tuỳ theo cấu hình và theo yêu cầu của bạn.
     + Sau khi cài đặt xong tiến hành khởi động lại Elasticsearch với “**sudo systemctl restart elasticsearch**” hoặc “ **sudo systemctl start elasticsearch** ”
  2. Tiến hành mở port Elasticsearch
     + Để Elasticsearch lắng nghe các ip ngoài ta setting file elasticsearch.yml “ với câu lệnh “ **sudo vi /etc/elasticsearch/elasticsearch.yml “**. Thêm **http.host: 0.0.0.0**  vào sau “ Set the bind address to a specific IP (IPv4 or IPv6) ” sau đó restart lại Elasicsearch.



Ảnh 1.10 Es cho phép ip ngoài connect

* + - Kiểm tra trạng thái tường lửa “ **sudo firewall-cmd --state**”.
    - Mở port 9200 **“ sudo firewall-cmd --zone=public --add-port=9200/tcp --permanent** ”
    - Reload lại firewall “ **sudo firewall-cmd --reload** “
    - Kiểm tra các port hoạt động “ **sudo firewall-cmd --zone=public --list-ports** “
    - Trước khi máy ngoài truy cập ES ta cần phải cho phép port từ máy ngoài truy cập. Máy tôi truy cập là máy windown nên ta cần phải setting tường lửa. Mở “ **Windows defender firewall… > Advanced settings > inbound Rule**”. Tiến hành thêm port 9200. Tham khảo thêm tại “ <https://kb.hostvn.net/hung-dn-m-cnginbound-firewall-tren-windows-10_724.html> “
    - Trên máy tính ngoài truy cập theo đường dẫn **http://<ip máy chủ>:9200** hoặc **http://domain:9200**



Ảnh 1.11 Truy cập Es từ máy ngoài

1. Elasticsearch

* So sánh Elasticsearch với relational database

|  |  |  |  |
| --- | --- | --- | --- |
| So sánh | Elasticsearch | relational database | NoSQL |
| Mô hình dữ liệu | Dựa trên mô hình dữ liệu phi cấu trúc (schema-less), sử dụng JSON để lưu trữ dữ liệu. Dữ liệu được tổ chức thành các tài liệu (documents) có cấu trúc linh hoạt. | Dựa trên mô hình dữ liệu có cấu trúc, sử dụng bảng, hàng và cột để tổ chức dữ liệu theo quy tắc và ràng buộc. | Có nhiều loại NoSQL, bao gồm hệ cơ sở dữ liệu key-value, cột gia đình (column family), tài liệu (document), đồ thời gian (time series), đồ đồ họa (graph), vv. Mỗi loại có mô hình dữ liệu riêng. |
| Tìm kiếm và truy vấn dữ liệu | Sử dụng Elasticsearch Query DSL để thực hiện truy vấn linh hoạt. Cung cấp tính năng tìm kiếm văn bản đầy đủ, tìm kiếm phù hợp, phân tích từ khóa và các tính năng tìm kiếm phức tạp khác. | Sử dụng SQL (Structured Query Language) để thực hiện các truy vấn có cấu trúc. Cung cấp tính năng tìm kiếm dựa trên các câu lệnh SELECT phức tạp. | Truy vấn và tìm kiếm dữ liệu có thể khác nhau tùy thuộc vào loại cơ sở dữ liệu NoSQL. Ví dụ, hệ cơ sở dữ liệu key-value chỉ hỗ trợ truy vấn theo khóa, trong khi hệ cơ sở dữ liệu đồ thời gian tập trung vào truy vấn dữ liệu theo thời gian. |
| Phân tán và mở rộng | Được thiết kế để hoạt động trong môi trường phân tán. Dữ liệu được chia nhỏ và phân phối trên nhiều nút trong một cụm Elasticsearch. Cung cấp khả năng mở rộng tuyến tính, cho phép xử lý tải lớn và tăng khả năng chịu lỗi. | Thường được triển khai trên một máy chủ duy nhất hoặc một cụm máy chủ có quy mô nhỏ. Cần các biện pháp phức tạp để mở rộng quy mô và xử lý tải cao. | Đa số hệ cơ sở dữ liệu NoSQL cung cấp khả năng mở rộng tốt hơn so với hệ cơ sở dữ liệu quan hệ. Có thể mở rộng theo chiều ngang (horizontal scaling) bằng cách thêm nút vào cụm hoặc theo chiều dọc (vertical scaling) bằng cách nâng cấp tài nguyên trên mỗi nút. |
| Tính năng phân tích và xử lý dữ liệu | Cung cấp các tính năng phân tích văn bản mạnh mẽ như tách từ (tokenization), lắp ghép từ (stemming), phân tích ngữ nghĩa (semantic analysis), tìm kiếm gần đồng nghĩa (fuzzy search) và nhiều khả năng tìm kiếm phức tạp khác. | Hỗ trợ tìm kiếm và phân tích cơ bản như tìm kiếm toàn văn, tìm kiếm theo từ khóa, và một số hàm xử lý văn bản đơn giản. | Tùy thuộc vào loại NoSQL, có thể cung cấp các tính năng phân tích dữ liệu phức tạp như xử lý đồ thời gian, phân tích đồ thị, tính toán song song, vv. |
| Ứng dụng | Thích hợp cho các ứng dụng tìm kiếm, phân tích và khám phá dữ liệu, nơi cần tìm kiếm nhanh, tính linh hoạt cao và khả năng xử lý dữ liệu phi cấu trúc. | Thích hợp cho các ứng dụng có dữ liệu có cấu trúc, cần tính nhất quán và kiểm soát ràng buộc. | Thích hợp cho các ứng dụng có yêu cầu mở rộng, tính nhất quán thấp, và khả năng xử lý dữ liệu không cấu trúc hoặc dữ liệu lớn. |

* Khi nào nên chon Elasticsearch
  + - Elasticsearch rất mạnh mẽ trong việc tìm kiếm và truy vấn dữ liệu. Nếu dự án của bạn đòi hỏi khả năng tìm kiếm nhanh chóng và phức tạp trên lượng lớn dữ liệu, Elasticsearch là lựa chọn tốt.
    - Hệ thống phân tích log: Elasticsearch được sử dụng phổ biến trong việc phân tích và xử lý log. Với tính năng tìm kiếm và phân tích dữ liệu thời gian thực, Elasticsearch có thể giúp bạn xử lý và hiểu rõ hơn về các log hệ thống, ứng dụng hoặc mạng.
    - Hệ thống đo lường và giám sát: Elasticsearch kết hợp với Kibana có thể tạo ra một hệ thống giám sát và đo lường mạnh mẽ. Nó cho phép bạn thu thập, lưu trữ và phân tích dữ liệu đo lường từ nhiều nguồn khác nhau, từ đó cung cấp thông tin quan trọng để theo dõi và điều chỉnh hệ thống.
    - Hệ thống gợi ý và đề xuất: Elasticsearch có khả năng tạo ra các gợi ý và đề xuất dựa trên dữ liệu tìm kiếm và phân tích. Nếu bạn cần xây dựng một ứng dụng có khả năng gợi ý và đề xuất cho người dùng, Elasticsearch có tính năng phù hợp.
    - Hệ thống tìm kiếm và phân tích dữ liệu văn bản: Elasticsearch được thiết kế đặc biệt để xử lý và phân tích dữ liệu văn bản. Nếu dự án của bạn liên quan đến xử lý và phân tích văn bản, Elasticsearch là một lựa chọn hợp lý.
    - Dự án có yêu cầu tìm kiếm và truy vấn dữ liệu phức tạp: Elasticsearch cung cấp Elasticsearch Query DSL cho phép bạn thực hiện truy vấn phức tạp và tìm kiếm linh hoạt trên dữ liệu. Nếu bạn cần xử lý truy vấn dữ liệu phức tạp, Elasticsearch có tính năng phù hợp.
* Thêm người dùng trong Elasticsearch
  + - Để thêm bật security trên elasticsearch ta cần thêm “ **xpack.security.enabled: true** ” cuối file **elasticsearch.yml.**
    - Để thiết lập tài khoản dùng câu lệnh: “**sudo /usr/share/elasticsearch/bin/elasticsearch-setup-passwords auto**”, “ **sudo /usr/share/elasticsearch/bin/elasticsearch-setup-passwords random**” bên cửa sổ sẽ thông báo 1 list user và password bạn cần lưu lại để login.
    - Cần tạo 1 tài khoản khác ta setting thêm “ **xpack.security.authc.api\_key.enabled: true** ”, khởi động lại elasticsearch, chạy lại setup password

<https://xuanthulab.net/cai-dat-elasticsearch-tren-centos.html>

<https://www.digitalocean.com/community/tutorials/how-to-install-and-configure-elasticsearch-on-centos-7>