TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
_____ * _____

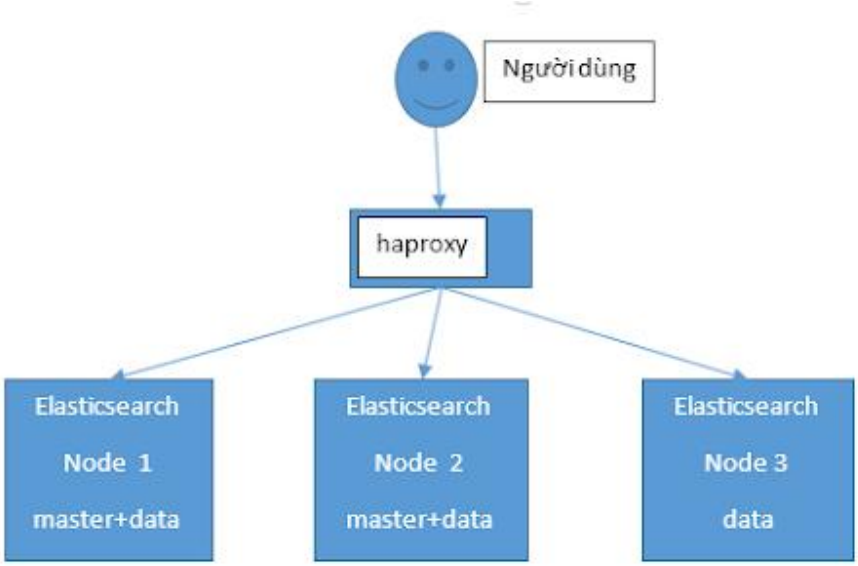
BÀI TẬP TRÊN LỚP

MÔN HỌC: LƯU TRỮ VÀ XỬ LÝ DỮ LIỆU LỚN

LAB 2: ELASTICSEARCH

NHÓM: GIỮA CHÚNG TA

Hà Nội, 11-2020



Sơ đồ cụm cluster Elasticsearch

Chúng ta sẽ cài đặt elasticsearch trên 3 máy chủ. Node 1 và node 2 làm master node, còn node 3 là data node.

I) Cài đặt ElasticSearch trên 3 máy:

- Bước 1: Luôn cập nhật hệ thống trước khi cài đặt

```
sudo apt-get update && apt-get upgrade
```

- Bước 2: Cài đặt JDK

```
sudo apt-get install default-jdk
```

Sau khi cài JDK xong, chúng ta có thể kiểm tra xem đã cài đặt thành công hay chưa bằng cách kiểm tra phiên bản JDK vừa cài bằng lệnh:

java -version

```
ly@ly:~$ java -version
openjdk version "11.0.9.1" 2020-11-04
OpenJDK Runtime Environment (build 11.0.9.1+1-Ubuntu-0ubuntu1.18.04)
OpenJDK 64-Bit Server VM (build 11.0.9.1+1-Ubuntu-0ubuntu1.18.04, mixed mode, sh
aring)
```

Nếu hiển thị như trong ảnh tức là đã cài đặt thành công.

- Bước 3: Cài đặt Elasticsearch

Chạy lần lượt các lệnh sau:

```
sudo apt-get install apt-transport-https

wget -qO - https://artifacts.elastic.co/GPG-KEY-
elasticsearch | sudo apt-key add -

sudo add-apt-repository "deb
https://artifacts.elastic.co/packages/5.x/apt stable main"

sudo apt-get update

sudo apt-get install elasticsearch
```

- Bước 4: Config

Mở file `elasticsearch.yml` bằng lệnh:

```
sudo nano /etc/elasticsearch/elasticsearch.yml
```

Chỉnh sửa như sau:

```
network.host: localhost

http.port: 9200
```

- Bước 5: Khởi động `Elasticsearch`

Chạy lần lượt các lệnh:

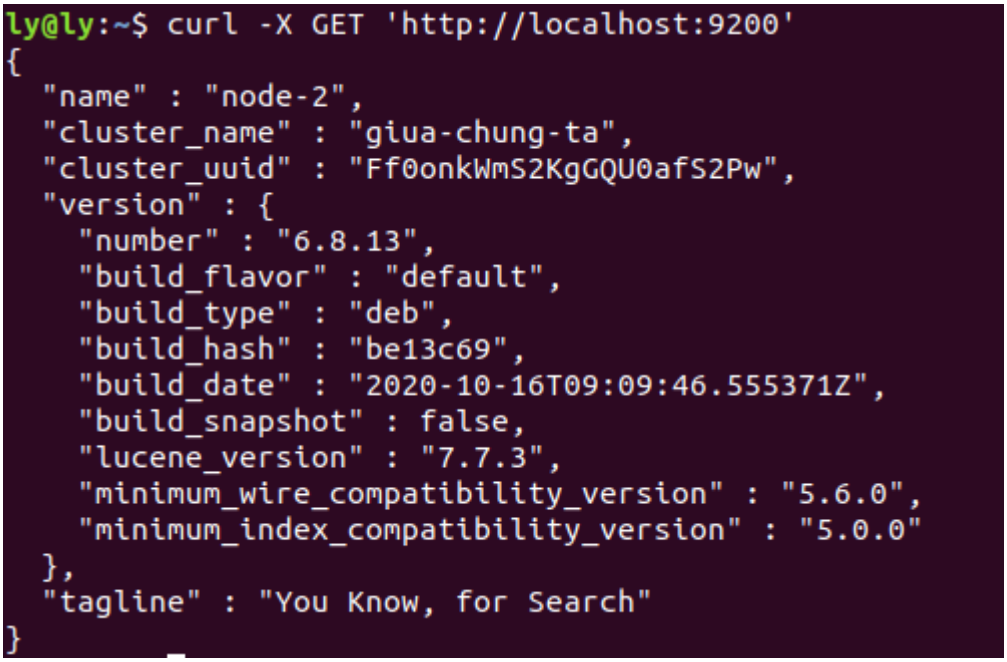
```
sudo /bin/systemctl enable elasticsearch.service

sudo systemctl start elasticsearch.service
```

Sau đó, ta có thể kiểm tra xem Elasticsearch đã được cài thành công trên máy hay chưa, nếu có hiển thị như ảnh dưới đây tức ta đã cài đặt thành công:

Sử dụng lệnh:

```
curl -X GET 'localhost:9200'
```



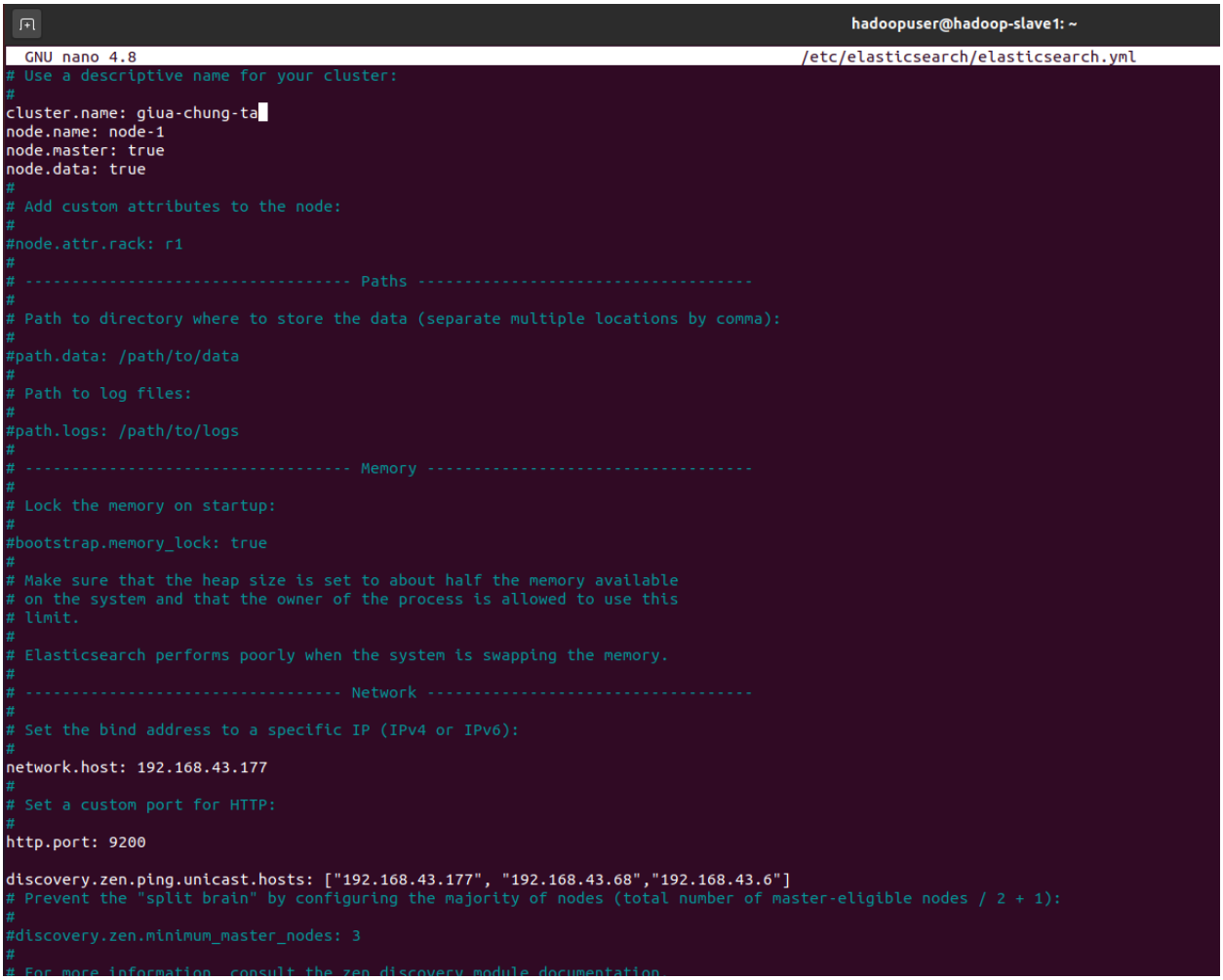
Làm lần lượt tất cả các bước trên cho cả 3 máy.

II) Tạo cụm Elasticsearch:

Chúng ta có 3 node: Node 1 và node 2 làm master node, node 3 làm data node. Đối với mỗi node, chúng ta đều cần mở tệp cấu hình Elasticsearch bằng câu lệnh sau:

```
sudo nano /etc/elasticsearch/elasticsearch.yml
```

Đối với node 1, chúng ta cấu hình như sau:



Node 2 cấu hình như sau:

```
  GNU nano 2.9.3                                    /etc/elasticsearch/elasticsearch.yml


# ---------------------------------- Cluster -----------------------------------
cluster.name: giua-chung-ta
#
# ------------------------------------ Node ------------------------------------
node.name: node-2
node.master: true
node.data: true
# ---------------------------------- Paths -------------------------------------
#
# Path to directory where to store the data (separate multiple locations by comma):
#
path.data: /var/lib/elasticsearch
#
# Path to log files:
#
path.logs: /var/log/elasticsearch
#
# ---------------------------------- Network -----------------------------------
network.host: 192.168.43.68
#
# Set a custom port for HTTP:
#
http.port: 9200
# --------------------------------- Discovery ----------------------------------
discovery.zen.ping.unicast.hosts: ["192.168.43.177", "192.168.43.68","192.168.43.6"]
```

Node 3:

```
#
cluster.name: giua-chung-ta
#
# Use a descriptive name for the node:
#
node.name: node-3
node.master: false
node.data: true
#
# Add custom attributes to the node:
#
#node.attr.rack: r1
#
# ---------------------------------- Paths ------------------------------------
#
# Path to directory where to store the data (separate multiple locations by comma):
#
path.data: /var/lib/elasticsearch
#
# Path to log files:
#
path.logs: /var/log/elasticsearch
# ---------------------------------- Network -----------------------------------#
# Set the bind address to a specific IP (IPv4 or IPv6):
#
network.host: 192.168.43.6
#
# Set a custom port for HTTP:
#
http.port: 9200
#
# For more information, consult the network module documentation.
#
# --------------------------------- Discovery ----------------------------------
discovery.zen.ping.unicast.hosts: ["192.168.43.177", "192.168.43.68","192.168.43.6"]
# For more information, consult the zen discovery module documentation.
#
# ---------------------------------- Gateway -----------------------------------
#
```

cluster-name: tên của cụm cluster, cấu hình cluster-name ở cả 3 máy phải giống nhau
network-host: điền địa chỉ IP của máy.

Sau đó, tại mỗi node, chúng ta chạy lệnh sau:

```
sudo service elasticsearch start
```

Nếu tất cả mọi thứ đã được cấu hình chính xác, cụm Elaticsearch của bạn sẽ hoạt động. Để xác minh mọi thứ đang hoạt động như mong đợi, hãy truy vấn Elaticsearch từ bất kỳ node nào bằng lệnh dưới đây:

```
curl -XGET 'http://[địa chỉ IP của
bạn]:9200/_cluster/state?pretty'
```

Kết quả thu được sẽ là data JSON giống như hình dưới:

Not secure | 192.168.43.177:9200/_nodes/node-2/stats?pretty

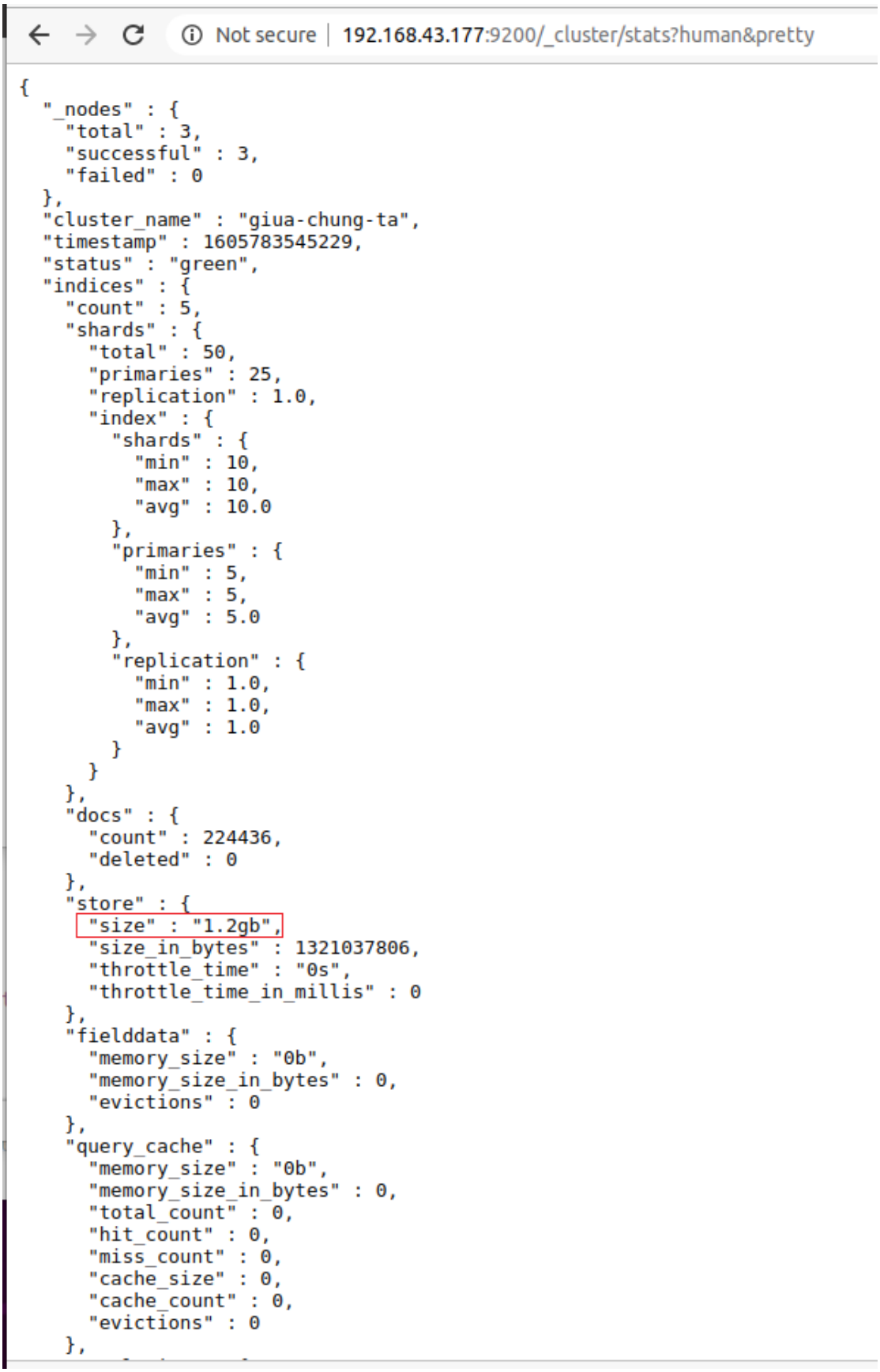

```
{
  "_nodes" : {
    "total" : 1,
    "successful" : 1,
    "failed" : 0
  },
  "cluster_name" : "giua-chung-ta",
  "nodes" : {
    "hoPXPZ__TGKBg9NPExX1LQ" : {
      "timestamp" : 1605783473290,
      "name" : "node-2",
      "transport_address" : "192.168.43.68:9300",
      "host" : "192.168.43.68",
      "ip" : "192.168.43.68:9300",
      "roles" : [
        "master",
        "data",
        "ingest"
      ],
      "attributes" : {
        "ml.machine_memory" : "16720945152",
        "ml.max_open_jobs" : "20",
        "xpack.installed" : "true",
        "ml.enabled" : "true"
      },
      "indices" : {
        "docs" : {
          "count" : 172045,
          "deleted" : 0
        },
        "store" : {
          "size_in_bytes" : 505685479,
          "throttle_time_in_millis" : 0
        },
        "indexing" : {
          "index_total" : 171657,
          "index_time_in_millis" : 118168,
          "index_current" : 0,
          "index_failed" : 0,
          "delete_total" : 0,
          "delete_time_in_millis" : 0,
          "delete_current" : 0,
          "noop_update_total" : 0,
          "is_throttled" : false,
          "throttle_time_in_millis" : 0
        },
        "get" : {
          "total" : 0,
          "time_in_millis" : 0,
          "exists_total" : 0,
          "exists_time_in_millis" : 0,
          "missing_total" : 0,
          "missing_time_in_millis" : 0,
          "current" : 0
        },
        "search" : {
          "open_contexts" : 0,
          "query_total" : 0,
```

III) Upload 1GB data lên Elasticsearch:

Chúng ta chuẩn bị 1GB file dữ liệu JSON.
Sau đó, `cd` đến thư mục lưu dữ liệu cần đẩy lên Elasticsearch.
Sau đó sử dụng câu lệnh sau để đưa dữ liệu vào Elasticsearch:

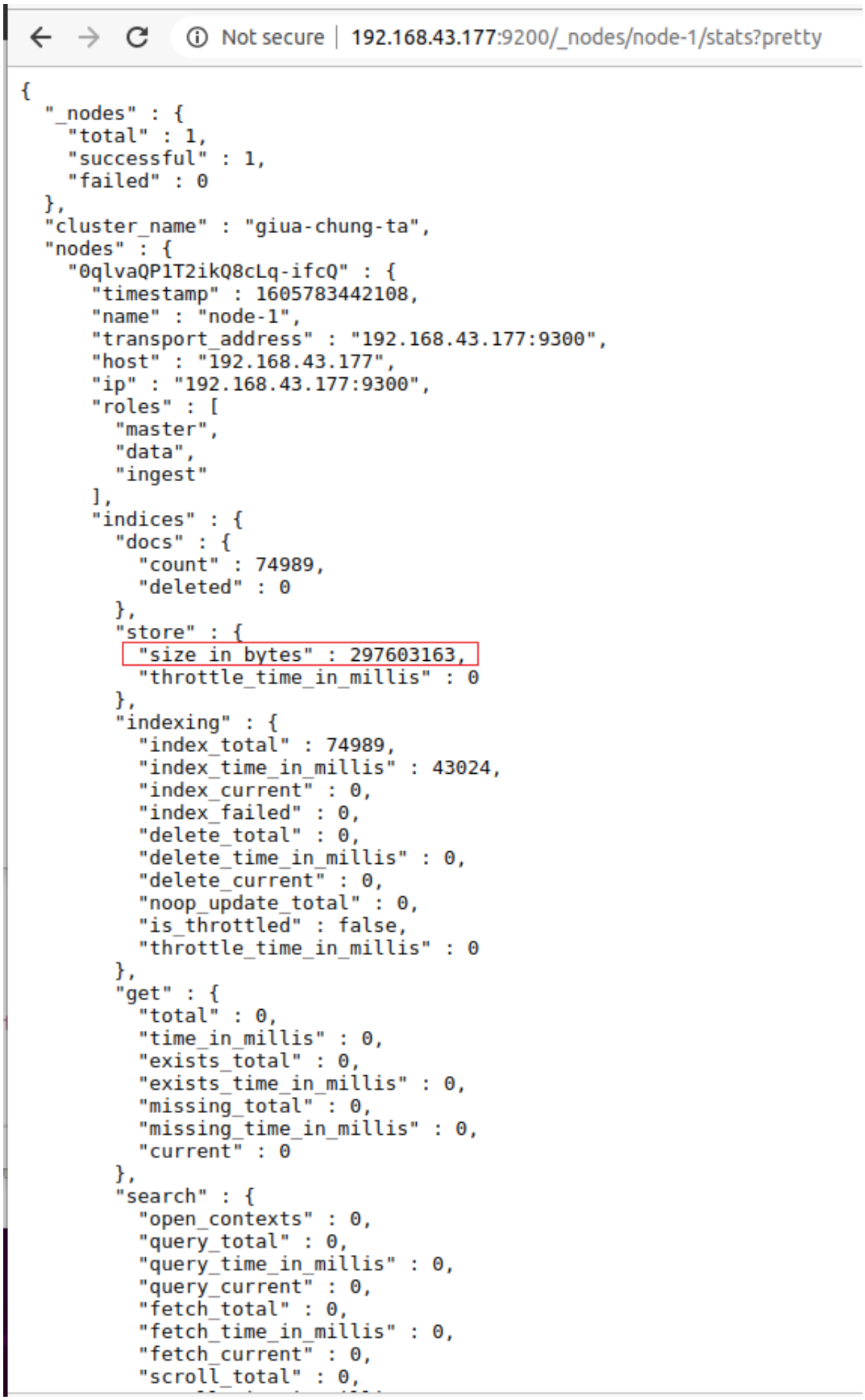
```
curl -H 'Content-Type: application/x-ndjson' -XPOST '[địa chỉ IP của bạn]:9200/[thư mục lưu trữ]/doc/_bulk?pretty' --data-binary @[tên file]
```

Minh chứng đã upload 1GB dữ liệu:

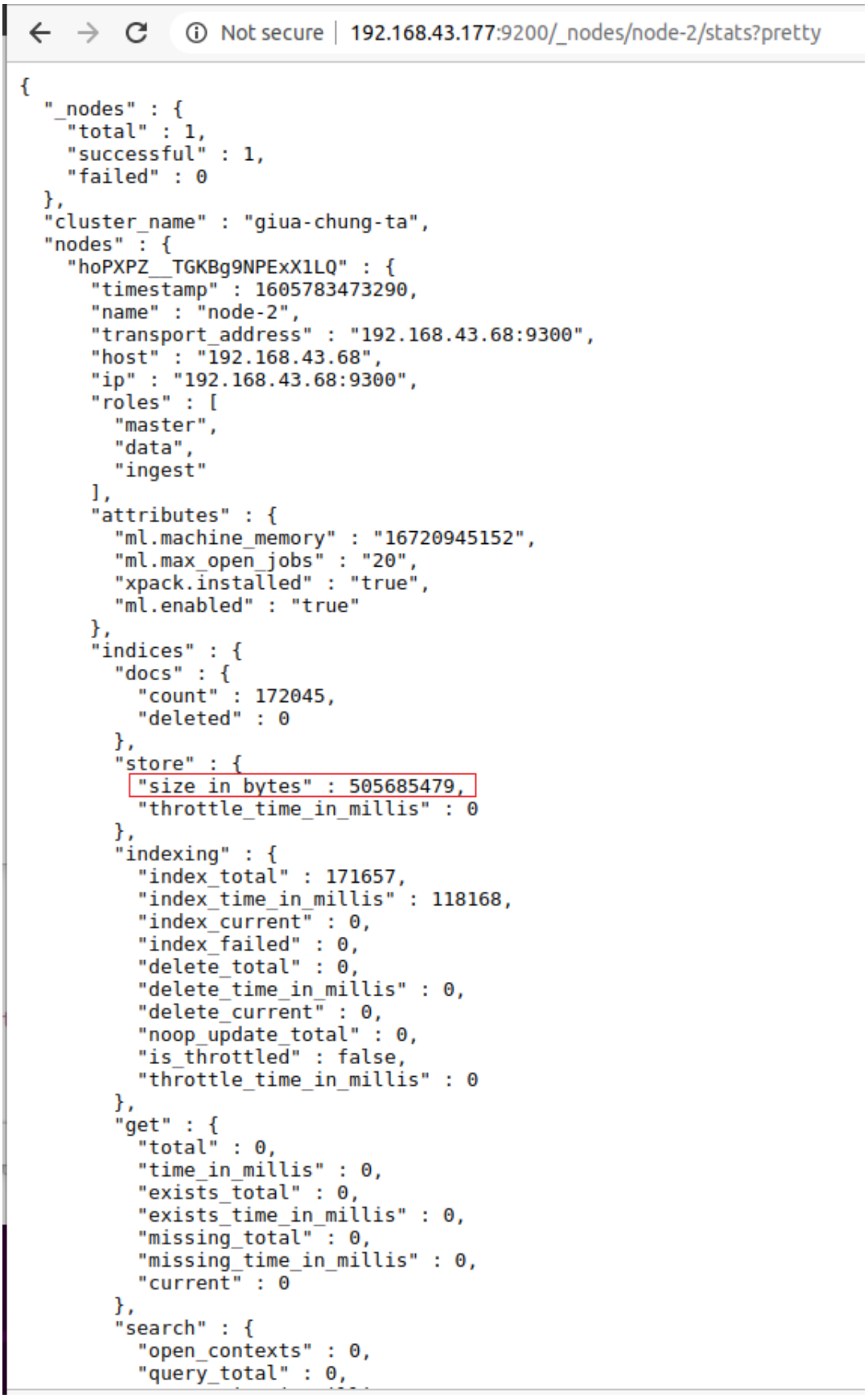

Not secure | 192.168.43.177:9200/_cluster/stats?human&pretty
{
"_nodes" : {
"total" : 3,
"successful" : 3,
"failed" : 0
},
"cluster_name" : "giua-chung-ta",
"timestamp" : 1605783545229,
"status" : "green",
"indices" : {
"count" : 5,
"shards" : {
"total" : 50,
"primaries" : 25,
"replication" : 1.0,
"index" : {
"shards" : {
"min" : 10,
"max" : 10,
"avg" : 10.0
},
"primaries" : {
"min" : 5,
"max" : 5,
"avg" : 5.0
},
"replication" : {
"min" : 1.0,
"max" : 1.0,
"avg" : 1.0
}
}
},
"docs" : {
"count" : 224436,
"deleted" : 0
},
"store" : {
"size" : "1.2gb",
"size_in_bytes" : 1321037806,
"throttle_time" : "0s",
"throttle_time_in_millis" : 0
},
"fielddata" : {
"memory_size" : "0b",
"memory_size_in_bytes" : 0,
"evictions" : 0
},
"query_cache" : {
"memory_size" : "0b",
"memory_size_in_bytes" : 0,
"total_count" : 0,
"hit_count" : 0,
"miss_count" : 0,
"cache_size" : 0,
"cache_count" : 0,
"evictions" : 0
},

Minh chứng dữ liệu được lưu phân tán:

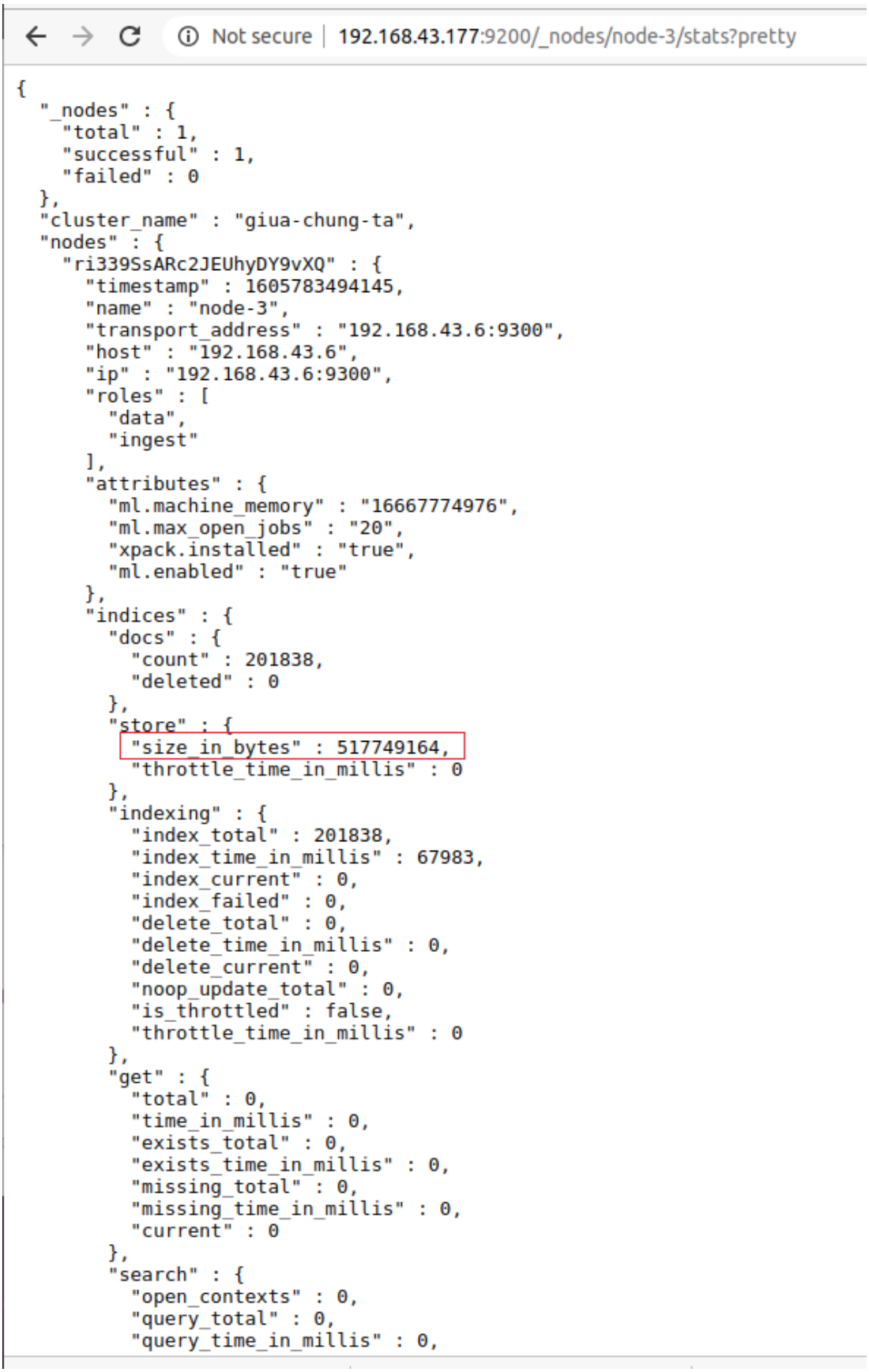
Trên node 1:


Not secure | 192.168.43.177:9200/_nodes/node-1/stats?pretty
{
"_nodes" : {
"total" : 1,
"successful" : 1,
"failed" : 0
},
"cluster_name" : "giua-chung-ta",
"nodes" : {
"0qlvaQP1T2ikQ8cLq-ifcQ" : {
"timestamp" : 1605783442108,
"name" : "node-1",
"transport_address" : "192.168.43.177:9300",
"host" : "192.168.43.177",
"ip" : "192.168.43.177:9300",
"roles" : [
"master",
"data",
"ingest"
],
"indices" : {
"docs" : {
"count" : 74989,
"deleted" : 0
},
"store" : {
"size_in_bytes" : 297603163,
"throttle_time_in_millis" : 0
},
"indexing" : {
"index_total" : 74989,
"index_time_in_millis" : 43024,
"index_current" : 0,
"index_failed" : 0,
"delete_total" : 0,
"delete_time_in_millis" : 0,
"delete_current" : 0,
"noop_update_total" : 0,
"is_throttled" : false,
"throttle_time_in_millis" : 0
},
"get" : {
"total" : 0,
"time_in_millis" : 0,
"exists_total" : 0,
"exists_time_in_millis" : 0,
"missing_total" : 0,
"missing_time_in_millis" : 0,
"current" : 0
},
"search" : {
"open_contexts" : 0,
"query_total" : 0,
"query_time_in_millis" : 0,
"query_current" : 0,
"fetch_total" : 0,
"fetch_time_in_millis" : 0,
"fetch_current" : 0,
"scroll_total" : 0,

Trên node 2:


Not secure | 192.168.43.177:9200/_nodes/node-2/stats?pretty
{
"_nodes" : {
"total" : 1,
"successful" : 1,
"failed" : 0
},
"cluster_name" : "giua-chung-ta",
"nodes" : {
"hoPXPZ__TGKBg9NPExX1LQ" : {
"timestamp" : 1605783473290,
"name" : "node-2",
"transport_address" : "192.168.43.68:9300",
"host" : "192.168.43.68",
"ip" : "192.168.43.68:9300",
"roles" : [
"master",
"data",
"ingest"
],
"attributes" : {
"ml.machine_memory" : "16720945152",
"ml.max_open_jobs" : "20",
"xpack.installed" : "true",
"ml.enabled" : "true"
},
"indices" : {
"docs" : {
"count" : 172045,
"deleted" : 0
},
"store" : {
"size_in_bytes" : 505685479,
"throttle_time_in_millis" : 0
},
"indexing" : {
"index_total" : 171657,
"index_time_in_millis" : 118168,
"index_current" : 0,
"index_failed" : 0,
"delete_total" : 0,
"delete_time_in_millis" : 0,
"delete_current" : 0,
"noop_update_total" : 0,
"is_throttled" : false,
"throttle_time_in_millis" : 0
},
"get" : {
"total" : 0,
"time_in_millis" : 0,
"exists_total" : 0,
"exists_time_in_millis" : 0,
"missing_total" : 0,
"missing_time_in_millis" : 0,
"current" : 0
},
"search" : {
"open_contexts" : 0,
"query_total" : 0,

Trên node 3:


Not secure | 192.168.43.177:9200/_nodes/node-3/stats?pretty
{
"_nodes" : {
"total" : 1,
"successful" : 1,
"failed" : 0
},
"cluster_name" : "giua-chung-ta",
"nodes" : {
"ri339SsARc2JEUhyDY9vXQ" : {
"timestamp" : 1605783494145,
"name" : "node-3",
"transport_address" : "192.168.43.6:9300",
"host" : "192.168.43.6",
"ip" : "192.168.43.6:9300",
"roles" : [
"data",
"ingest"
],
"attributes" : {
"ml.machine_memory" : "16667774976",
"ml.max_open_jobs" : "20",
"xpack.installed" : "true",
"ml.enabled" : "true"
},
"indices" : {
"docs" : {
"count" : 201838,
"deleted" : 0
},
"store" : {
"size_in_bytes" : 517749164,
"throttle_time_in_millis" : 0
},
"indexing" : {
"index_total" : 201838,
"index_time_in_millis" : 67983,
"index_current" : 0,
"index_failed" : 0,
"delete_total" : 0,
"delete_time_in_millis" : 0,
"delete_current" : 0,
"noop_update_total" : 0,
"is_throttled" : false,
"throttle_time_in_millis" : 0
},
"get" : {
"total" : 0,
"time_in_millis" : 0,
"exists_total" : 0,
"exists_time_in_millis" : 0,
"missing_total" : 0,
"missing_time_in_millis" : 0,
"current" : 0
},
"search" : {
"open_contexts" : 0,
"query_total" : 0,
"query_time_in_millis" : 0,