**BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ QUỐC PHÒNG**

HỌC VIỆN KỸ THUẬT QUÂN SỰ

**NGHIÊN CỨU SEARCH ENGINE ELASTICSEARCH VÀ HỆ SINH THÁI ELK STACK**

**Chuyên ngành: Công nghệ phần mềm**

ĐỀ CƯƠNG ĐỒ ÁN MÔN HỌC

**Hà Nội - Năm 2019**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO BỘ QUỐC PHÒNG**

HỌC VIỆN KỸ THUẬT QUÂN SỰ

**ĐỀ CƯƠNG ĐỒ ÁN MÔN HỌC**

Chuyên ngành: Công nghệ phần mềm

Ngày giao đồ án: .......................................................................

Ngày nộp đồ án: ........................................................................

Tên đề tài: **Nghiên cứu search engine Elasticsearch và hệ sinh thái ELK stack**

**Học viên thực hiện**: Trần Văn Hiếu

**Lớp**: KTPM14

**Cán bộ hướng dẫn**:

**Họ và tên**: Vi Bảo Ngọc

**Cấp bậc**:

**Học hàm, học vị**:

**Đơn vị**:

Hà Nội - Năm 2019

# ĐỀ CƯƠNG ĐỒ ÁN MÔN HỌC

Tên đề tài: Nghiên cứu search engine Elasticsearch và hệ sinh thái ELK stack

Chuyên ngành: Công nghệ phần mềm

Thời gian thực hiện: 4 tuần (từ 28/08 – 25/09/2018)

## I. Cơ sở khoa học và tính thực tiễn của đề tài

1. **Cơ sở khoa học**

Nhu cầu tìm kiếm trong tập dữ liệu lớn, tốc độ và chất lượng kết quả tìm kiếm được đề cao trong nhiều năm trở lại đây. Vì thế hàng loạt các search engine ra đời để giải quyết vấn đề này như: Elasticsearch, Sphinxsearch, Solr, … Các bài toán được đặt ra cho các search engine này là khả năng lưu trữ, tốc độ truy xuất làm sao cho ưu việt nhất.

Gần đây có nhiều mô hình giải pháp hướng dịch vụ ra đời. Điều đặt ra ở đây là khi ứng dụng đang release xảy ra lỗi mà chúng ta không biết lỗi này nằm ở service nào khi có quá nhiều service được chạy phục vụ ứng dụng. Vì vậy người ta cần đến các ứng dụng ghi log tập trung. Điều này góp mặt trong sự ra đời của hệ sinh thái elk stack để phục vụ quản lý log, tìm kiếm, thống kê…

1. **Tính thực tiễn**

Hiện này nhu cầu thương mai điện tử gia tăng với tần suất cao, hàng ngày có hàng triệu đơn hàng, hàng triệu sản phẩm được trưng bán đối với những trang thương mại lớn. Để giải quyết bài toán về tốc độ cho 1 trang điện tử như vậy quả lan giải. Người ta đã nghĩ đến cách sử dụng những search engine thay vì sử dụng các database thông thường để lưu trữ những dữ liệu thường xuyên được sử dụng. Database thông thường có đánh index, chia partition cũng chỉ cải thiện phần nào tốc độ truy vấn xong quá nhiều transaction phục vụ cho mục đích tìm kiếm cũng sẽ ảnh hưởng rất nhiều đến hệ thống.

Để release 1 trang web với số lượng người dùng đồng thời như facebook, zalo, tiki, lazada, … Thì các ứng dụng bắt buộc phải sử dụng đến dịch vụ phân tán. Điển hình với cơ chế loadbalancer 1 service chúng ta có thể đặt ở nhiều máy khác nhau để share tải cho nhau. Như vậy khi 1 node nào đó xảy ra sự cố chúng ta phải lần mò hết tất cả các server lấy file log ra rồi tìm xem log nằm ở file nào. Làm như thế thật tốn công sức. Người ta đã nghĩ đến tạo ra 1 ứng dụng quản lý ghi log tập trung.

## II. Mục tiêu của đề tài

### **Mục tiêu tổng quát:**

- Đề tài giúp tìm hiểu về search engine elasticsearch cách tiếp cận và xử lý bài toán lưu trữ với dữ liệu lớn truy vấn nhanh của elasticsearch cách triển khai nó trên môi trường phân tán. Hiểu về kiến trúc Master/Slave của nó.

- Sử dụng phối hợp giữa elasticsearch, kibana, logstash để quản lý log, phân tích dữ liệu lớn thu thập về elasticsearch bằng kibana. Sử dụng restfull từ api, kết hợp giữa logtash với log4j, log4net,… . Tích hợp hệ sinh thái này vào các ứng dụng có sẵn.

**2. Mục tiêu chi tiết**

- Đề tài tập trung vào phân tích tính ưu việt của elasticsearch. Các xử lý bài toán lưu trữ của nó. Khả năng tùy biến của nó khi cung cấp các api restfull. Tốc độ đánh index của nó có gì khác biệt. Hiểu được các cấu trúc cấu thành nên elasticsearch.

- Nắm được cách vận hành hệ sinh thái ELK trong các ứng dụng thực tiễn và vai trò của từng thành phần trong hệ sinh thái đó.

## III. Phương pháp nghiên cứu

1. **Về lý thuyết:**

Thu thập thông tin thông qua tài liệu của tổ chức sáng lập và tài liệu liên quan. Phân tích nguồn tài liệu từ đó để hiểu được những khái niệm elasticsearch là gì, kibana, logstash có vai trò gì trong hệ sinh thái, Thế nào là cluster, index, node, shard trong elasticsearch.

Dựa trên sự so sánh với các truy vấn like search của database so sánh với tốc độ truy xuất của elasticsearch để có đánh giá về search engine, Với các ứng dụng chưa áp dụng ghi log tập trung thì công tác tìm lỗi với các ứng dụng release thật sự khó khăn thấy được tầm quan trọng của ghi log tập trung.

**Elasticsearch** là một trong số engine điển hình và có hiệu quả giải quyết vấn đề này cao. Với search engine là elasticsearch, hệ sinh thái elk stack ứng dụng ghi log tập trung và phân tích dữ liệu.

**Full text searching** và **Inverted Index**: Full text searching là cách tự nhiên nhất để tìm kiếm thông tin, Inverted Index là kĩ thuật index theo đơn vị term (thay vì index theo đơn vị row giống như mysql). Cụ thể hơn, Inverted Index là một cấu trúc dữ liệu, nhằm mục đích mapping giữa term, và các document chứa term đó.

**Kiến trúc client-server**: Là một kiến trúc hướng dịch vụ cụ thể elasticsearch sử dụng REST API để tương tác với client.

Chúng ta hãy nhìn vào những khái niệm chính của ElasticSearch:

**Cluster**: Một tập hợp Nodes (servers) chứa tất cả các dữ liệu.

**Node**: Một server duy nhất chứa một số dữ liệu và tham gia vào cluster’s indexing and querying.

**Index**: Hãy quên SQL Indexes đi. Mỗi ES Index là 1 tập hợp các documents.

**Shards**: Tập con các documents của 1 Index. Một Index có thể được chia thành nhiều shard.

**Type**: Một định nghĩa về schema of a Document bên trong một Index (Index có thể có nhiều type).

**Document**: Một JSON object với một số dữ liệu. Đây là basic information unit trong ES.

Kết quả nghiên cứu trong đề tài này có thể được sử dụng để: Triển khai, tích hợp vào những hệ thống đã tồn tại có nhu cầu lớn về tìm kiếm nâng cao, ghi log tập trung.

1. **Về thực nghiệm:**

Tiến hành cài đặt elastich search trên single node và multi node. Cài đặt logtash, kibana trên trên 1 node. Cấu hình các file config của log4j, log4net trỏ ghi log thông qua logtash trỏ về elasticsearch master. Tiến hành đánh index cho elasticsearch từ database đối với bảng dữ liệu được truy vấn tìm kiếm với số lượng lớn và độ chính xác cao (fulltext search).

Sử dụng các API có sẵn của elastich search để thực hiện thao tác tìm kiếm. Dựa vào các plugin có sẵn của kibana để phân tích đánh giá tình trạng truy vấn dữ liệu của người dùng. Phân tích đánh giá tần xuất xảy ra log error của hệ thống để khắc phục.

## IV. Nội dung nghiên cứu

* + Nội dung luận văn gồm một số vấn đề chính sau:

**LỜI GIỚI THIỆU**

**ĐẶT VẤN ĐỀ**

**Chương 1: Tổng quan**

* 1. Khái niệm về Search engine và Elasticsearch, ELK stack
     1. Search engine
     2. Elasticsearch
     3. ELK stack
  2. Các nhu cầu cần sử dụng đến Elasticsearch và ELK stack

**Chương 2: Vấn đề**

    2.1. Khác nhau giữa Elastichsearch và Database thông thường

2.1.1. Hạn chế của Database thông thường

2.1.2. Ưu nhược điểm của Elasticsearch so với Database

    2.2. Sự cần thiết của Elasticsearch và ELK stack

    2.3. Một số hệ thống sử dụng ELK stack

**Chương 3: Phân tích**

    3.1. Tìm hiểu về mô hình Master/Slave

    3.2. ELK stack

3.2.1. Cài đặt ELK

3.2.2. Tìm hiểu về quá trình ghi log của logtash lên elasticsearch

3.2.3. Khả năng phân tích dữ liệu của kibana

    3.3. Elasticsearch

3.3.1. Tìm hiểu về Indexing

3.3.2. Cài đặt Elasticsearch trên cụm cluster

    3.4. Một số hệ thống search engine khác liên quan

3.4.1. Sphinxsearch

3.4.2. Apache Solr

    3.6. Tích hợp với các api có sẵn của Elasticsearch

    3.7. Kết quả đạt được

**Chương 4: Kết luận**

## V. Tài liệu tham khảo:

[1] Elasticsearch BV, (2019, Sep 13), *Elasticsearch Reference* [Online],Available: <https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html>

[2] Elasticsearch BV, (2019, Sep 13), *Kibana Guide* [Online], Available: <https://www.elastic.co/guide/en/kibana/7.3/index.html>

[3] Elasticsearch BV, (2019, Sep 13), *Logstash Reference* [Online], Available: <https://www.elastic.co/guide/en/logstash/7.3/index.html>

[4] C. Bhadane, H. A. Mody, D. U. Shah, P. R. Sheth, (2014 Jan ), *Use of Elastic Search for Intelligent Algorithms to Ease the Healthcare Industry* [Online], Available: <http://www.ijsce.org/wp-content/uploads/papers/v3i6/F2013013614.pdf>

## VI. Dự kiến kế hoạch thực hiện:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Nội dung công việc** | **Thời gian thực hiện** | **Địa điểm thực hiện** | **Ghi chú** |
| 1 | Tìm hiểu về elasticsearch | 1 tuần | Tại nhà |  |
| 2 | Tìm hiểu về kibana, logtash | 3 ngày | Tại nhà |  |
| 3 | Tìm hiểu về kiến trúc ELK stack và ứng dụng | 2 tuần | Tại nhà |  |
| 4 | Cài đặt Elasticsearch multi node | 1 tuần | Tại nhà |  |
| 5 | Tích hợp ghi log vào các ứng dụng có sẵn | 1 tuần | Tại nhà |  |
| 6 | Ứng dụng api search của elasticsearch | 2 ngày | Tại nhà |  |
| 7 |  |  |  |  |

## VII. Các cơ quan, đơn vị cần liên hệ:

- Bộ môn Khoa học máy tính khoa Công nghệ Thông tin - Học viện Kỹ thuật Quân sự.

## VIII. Kinh phí thực hiện đề tài:

Ngày tháng năm 2019

**Chủ nhiệm bộ môn Người lập đề cương**

**Chủ nhiệm khoa Giáo viên hướng dẫn**