

УДК 004.415.25

РАЗРАБОТКА СИСТЕМЫ ПОИСКА ПО ИСТОРИКО-КУЛЬТУРОЛОГИЧЕСКИМ ДАННЫМ НА БАЗЕ ELASTICSEARCH И MONGODB

Мельников Н.П. (Университет ИТМО), Мороз И.О. (Университет ИТМО)

Научный руководитель – к. культурологии, ординарный доцент, Пучковская А.А.
(Университет ИТМО)

В работе представлены результаты проектирования и разработки информационной системы для решения проблемы поиска по историко-культурологическим данным из базы данных под управлением MongoDB с использованием Elasticsearch. Затрагиваются также аспекты развёртывания этих компонентов и внутренней работы компонентов Elasticsearch.

Введение. На сегодняшний день проблема настройки интеллектуального поиска в объёмных фактологических базах данных, содержащих информацию историко-культурологического характера, остаётся актуальной. Предлагаемое решение было разработано для базы краеведческих данных, в которой хранится информация о локациях, персонах и связях между локациями и персонами, а также некоторые другие исторические сведения. Для удобства работы с данными был разработан пользовательский интерфейс для просмотра и редактирования информации из базы данных. Предлагаемое решение даст возможность продвинутого поиска по хранимым данным на основе различных критериев и фильтров.

Основная часть. Для решения проблемы был использован продукт Elasticsearch компании Elastic NV. Elasticsearch это распределённая система для поиска, анализа и хранения данных с доступом по JSON REST API. Он входит в стек продуктов ELK (Elasticsearch + Logstash + Kibana) — комплексное решение для сбора, анализа и визуализации данных.

Для развёртывания MongoDB, Elasticsearch и Kibana использовались Docker и docker-compose. Был написан набор конфигурационных файлов для простого запуска и настройки этих решений как на локальной машине, так и на удалённом сервере.

Одной из проблем, возникших в процессе решения, была синхронизация данных между MongoDB и Elasticsearch, так как основное хранилище информации находится под управлением MongoDB, а Elasticsearch используется исключительно в качестве поисковой системы. Для синхронизации данных использовался monstache — open-source решение на языке программирования Go, которое работает в качестве фонового потока. Он следит за изменениями данных в MongoDB и направляет поток изменений в Elasticsearch.

Для эффективного поиска в Elasticsearch было необходимо правильно проиндексировать данные в соответствии с запросами пользователей. Основные требования при реализации поиска были следующими:

- Устойчивость к опечаткам при вводе запроса и возможность их автоматического исправления;
- Устойчивость к синонимам;
- Автодополнение при вводе запроса;
- Возможность фильтрации по определённому временному промежутку;
- Поиск по адресам: должен вернуть локацию по её адресу;
- Поиск по координатам: должен вернуть локацию по переданным координатам;
- Поиск персон по именам;
- Поиск персон и локаций по тегам (например, по запросу «серебряный век» показать персон, которые относятся к нему).

Для тестирования и отладки запросов в Elasticsearch использовался сервис Kibana. Данный сервис также предоставляет богатые возможности для анализа и визуализации данных. Доступ к нему могут иметь не только администраторы системы, но и обычные пользователи, наравне с интерфейсом администратора, для удобного анализа имеющихся данных и построения интерактивных графиков и информационных панелей на их основе.

Выводы.

В результате проделанной работы удалось построить продвинутую поисковую систему для поиска по историко-культурологическим данным. Были решены следующие проблемы:

- Синхронизация данных между MongoDB и Elasticsearch;
- Настройка Docker и docker-compose для работы с базой данных как локально, так и на сервере;
- Настройка поиска на базе Elasticsearch для обработки широкого спектра пользовательских поисковых запросов с применением различных фильтров и устойчивостью к опечаткам и ошибкам;
- Настройка Kibana для анализа и визуализации данных в Elasticsearch, а также для отладки запросов.

Мельников Н.П. (автор)

Подпись

Мороз И.О. (автор)

Подпись

Пучковская А.А. (научный руководитель)

Подпись