Використання Стеку ELK для Обробки Великих Даних

Ольга Заверуха
кафедра Прикладної математики
Харківський національний університет
радіоелектроніки
Харків, Україна
olha.zaverukha@nure.ua

Володимир Кобзєв
кафедра Прикладної математики
Харківський національний університет
радіоелектроніки
Харків, Україна
volodymyr.kobziev@nure.ua

Using the ELK-stack for Big Data Processing

Olha Zaverukha
Applied Mathematics Department
Kharkiv National University of Radio Elektronics
Kharkiv, Ukraine
olha.zaverukha@nure.ua

Volodymyr Kobziev
Applied Mathematics Department
Kharkiv National University of Radio Elektronics
Kharkiv, Ukraine
volodymyr.kobziev@nure.ua

Анотація—У даній статті розглядаються можливості ELK-стеку, що утворює систему зчитування, зберігання, швидкої обробки та якісної візуалізації великих даних. Узагальнено наводяться переваги стеку, приклад його використання для обробки та відображення туристичних даних.

Abstract—This article examines the capabilities of the ELK-stack, which creates a system for reading, storing, fast processing and high-quality visualization of large data. Generally, the advantages of the stack are given, an example of its use for processing and displaying tourist data

Ключові слова—Великі дані, стек, обробка, зберігання, візуалізація.

Keywords—Big Data, stack, processing, storage, visualization.

I. Вступ.

Величезні обсяги, складність, мінливість, слабка структурованість та різноманітність є характерними рисами даних, що підлягають кількісному та якісному аналізу і сприйнятному відображенню, у різних сферах людської діяльності протягом останніх десятиліть. Збір, зберігання та аналіз даних з вказаними особливостями являють собою доволі складну проблему, впоратися з якою можна тільки завдяки технологіям Big Data.

Однією з найпоширеніших є парадігма розподіленої обробки даних MapReduce, яка розрахована на участь у зберіганні, кількох етапах обробки та відображенні кластерів з багатьох комп'ютерів. Інший підхід базується на використанні ELK-стеку.

II. Загальна характеристика стеку ELK

Стек ELK створений у 2010 році як потужна система гнучкого пошуку, зберігання та візуалізації даних, що має три складові (Elasticsearch, Logstash та Kibana), наведені на Рис.1.

Рис. 1. Структура стеку ELK

Протягом останніх п'яти років вона стала дуже популярною серед аналітиків Великих даних завдяки своїм можливостям, простоті освоєння і застосування [1, 2]. Ця система здатна збирати та аналізувати десятки мільйонів даних за день, що саме й розуміють під поняттям Big Data.

Підсистема Logstash має можливості передати дані, вилучити їх з бази даних, з файлу, або з встановленого клієнту, що збирає логи, а також може перетворити їх на потрібний формат чи структуру та завантажити у будь-яке сховище даних. На доданок до цього, вона має змогу у будь-який час передати дані до підсистеми Elasticsearch.

За виконуваними функціями Logstash – це доволі простий інструмент, який передає дані з одного чи багатьох входів та виводить їх на один, чи декілька виходів. Входами можуть бути будь-які джерела даних, такі як файли журналів або логи. Після того як вхідні дані будуть зчитані, logstash аналізує їх за допомогою кодеків, таких як JSON [2].

Під час переміщення даних з джерела в сховище, фільтри Logstash аналізують кожну подію, ідентифікують назви полів для побудови структури та перетворюють їх у загальний формат для полегшення та прискорення аналізу. Підсистема Logstash динамічно трансформує та готує дані незалежно від формату або складності: розшифровує географічні координати, створює структуру з неструктурованих даних, анонімізує дані. Простота обробки даних не залежить від джерела, формату чи схеми даних.

Файл конфігурації Logstash має структуру, що наведена на Рис.2.

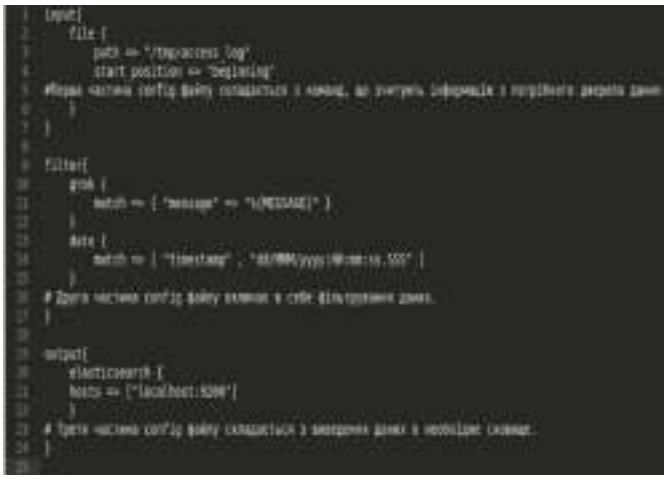

Рис. 2. Приклад файлу конфігурації для Logstash

Крім того, блок Logstash відстежує введені дані, які він обробляє. Таким чином, можна перезапустити його без шкоди дублювання даних. Хоча Logstash створений в Ruby, працює він дуже швидко. Пакетна версія запускається на JRuby, і вона використовує можливості JVM в потоці, даючи десятки потоків для паралельної обробки даних.

Блок Logstash може зберігати дані у блоці ElsaticSearch, який також запущений на машині, де працює Logstash. Підсиситема ElsaticSearch виконує роль сховища даних. Але це не база даних, яка буде сприймати все, що користувач забажає помістити до неї. Блок ElsaticSearch масштабує дані, коли це потрібно.

Стан Elasticsearch [3] наведений на Рис.3.

Рис. 3. Відображення стану Elasticsearch у браузері

Він відображається у один рядок і містить інформацію про статус, очікування інформації, зайнятий обсяг пам’яті, тощо.

За допомогою простої команди (/_cat/indeces) можна побачити наступну інформацію: стан індексу, статус, назву, об’єм, кількість рядків, розмір на диску, тощо (див. Рис. 4).

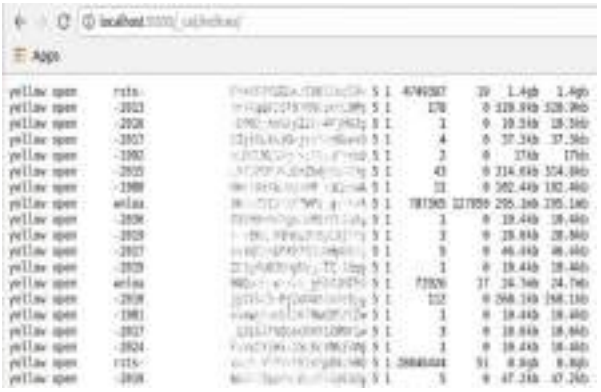

Рис. 4. Відображення індексів ElsaticSearch у браузері та інформації про них

Останнім кроком проведеного аналізу є відображення отриманих результатів, яке виконується підсистемою Kibana. Вона вже пов’язана з ElasticSearch, тому всі індекси вже знаходяться в ній, тож треба лише вибрати той, який у подальшій роботі необхідний для аналізу.

Kibana - єдиний компонент, який не резервується, так як вихід його з ладу не впливає на процеси збору і обробки інформації. Kibana має добрий інтерфейс для візуалізації величезної кількості даних і дозволяє краще зрозуміти дані через аналіз та візуалізацію, що можна створити за лічені хвилини. Незважаючи на простоту даних інструментів, вони дуже потужні. Наявність шаблонів розгортання дозволяє легко вибрати апаратні та архітектурні профілі, які найкраще відповідають потребам користувача.

Підсистема Kibana дозволяє будувати аналіз за такими напрямками, як геолокація, кількість, часові проміжки, тощо [4]. Окрім цього, всі зображення мають гарну графіку. Їх можна об’єднувати у інформаційні панелі dashboards, завдяки чому весь необхідний аналіз може бути зібраний на одній сторінці (див. Рис.5).

Ще однією безсумнівною перевагою стеку ELK є можливість автоматичного опрацювання нових даних без додаткового настроювання.

Стек ELK є системою, що відстежує інформацію у реальному часі. Тому після запуску усіх компонент, якщо джерело входу інформації буде поповнюватися новими даними, то усі компоненти будуть відстежувати їх, додавати у сховище та відображати на візуалізаціях без перезапуску їх самих. Це дуже зручно, якщо дані надходять постійно.

Рис. 5. Приклад панельної візуалізації за допомогою Kibana

III. Застосування ELK-стеку

Для наглядної демонстрації роботи ELK-стеку обрані дані про країни, які полюбляють відвідувати туристи. Рис. 6 свідчить, що кількість туристів, які прибули за один день до Індії, Таїланду, Австралії та Японії переважає показники інших країн.

Крім відвідуваності, також є статистика запитів в інтернеті про найпривабливіші країни для туристів. Кількість запитів на день переважає у Тайланду, про що свідчать дані наведені на рис. 7. Також важливо враховувати не тільки запити, але і реальні дані про країни та їх міста, які популярні серед туристів, про що гарно свідчить подвійна кругова діаграмія. На рис. 8 країни та міста умовно позначені числами.

Рис. 6. Візуалізація кількості відвідувань країн за один день

Рис. 7. . Кількість запитів в інтернеті щодо подорожей до країн за один день

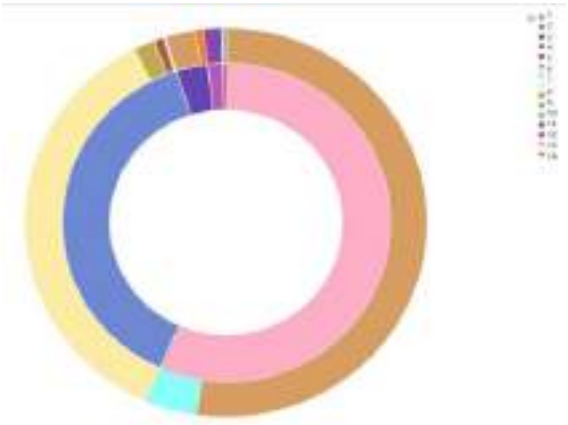

Рис. 8. Кругова діаграма популярності міст та країн

Швидкість роботи та візуалізації за допомогою ELK-стеку дійсно вражає. Аби завантажити дані та побудувати зрозумілі графіки, знадобилося 2 години. Притому, що кількість даних сягала приблизно 30 мільйонів.

ELK-стек дозволяє не тільки будувати кругові або стовпчасті діаграми, але й дає змогу чітко візуалізувати дані з географічними полями на детальній мапі світу.

Окрім яскравої та зрозумілої графіки, Kibana може виводити згруповані та відсортовані дані у таблицях, метриках або строках.

IV. Висновок

Таким чином, ELK-стек є доступним, високоякісним, надійним і простим у обслуговуванні інструментом, який надає можливість за лічені хвилини доставити, впорядкувати та візуалізувати Великі дані.

Дана система надає можливості, необхідні для створення комерційних звітів (унікальні підрахунки, послідовності переходів тощо).

ELK-стек добре захищений та швидко працює, що дуже важливо для систем обробки Великих даних.

Література References

[1] The Elastic Stack 6.5 [Online]. Available: https://www.elastic.co/

[2] Big data in minutes with the ELK Stack [Online]. Available: https://brewhouse.io/blog/2014/11/04/big-data-with-elk-stack.html

[3] Saurabh Chhajed “Learning ELK Stack”, 2015, Packt Publishing.

[4] Google Trends Analytics Data [Online]. Available: https://trends.google.ru/trends/?geo=RU