|  |  |
| --- | --- |
| **프로젝트 결과 보고서** | |
| 작 품 명 | ELK를 활용한 실시간 공공데이터 분석 |
| 팀 명 | 열혈분석 |
| 제 작 기 간 | 2019년 8월 12일 ~ 2019년 8월 16일 |
| 프로젝트 매니저 | 남명우 (인) |

|  |  |  |  |
| --- | --- | --- | --- |
| 조 원 | | | |
| No. | 성 명 | 연 락 처 | E-mail |
| 1 | 남명우 | 010-7601-0696 | Myungwoo0221@gmail.com |
| 2 | 심정현 | 010-7620-2146 | sim1152@naver.com |
| 3 | 김지훈 | 010-2475-0435 | rlawlgns0435@gmail.com |
| 4 | 안지원 | 010-7711-4523 | happytree213@naver.com |
| 5 | 장채연 | 010-2743-4833 | wkdcodus@gmail.com |

1. **Team 소개**

* **팀명**

교육이 끝난 후 함께 '열혈분식' 이라는 상호의 김밥 한 줄로 저녁을 때우며 프로젝트를 진행하여 '열혈분식' 이라는 단어가 인상 깊었고, 빅데이터 분석 과정에 맞게 '분식'을 '분석' 으로 바꿔서 '열혈분석'을 팀명으로 정하게 되었습니다. '열혈' 이라는 단어처럼 열렬한 정신과 격렬한 정열을 쏟아 데이터를 분석하는 프로젝트를 진행해봤습니다.

* **역할**

남명우 : 데이터 수집 / 데이터 정제 / 데이터 베이스 구축 / ELK연동 / 보고서 작성

심정현 : CentOS 설치 / MariaDB 설치 / 데이터 베이스 구축 / 보고서 작성

김지훈 : 데이터 베이스 구축 / 데이터 수집 / 데이터 정제 / ELK연동

안지원 : CentOS 설치 / MariaDB 설치 / Kibana 시각화 / 데이터 베이스 구축

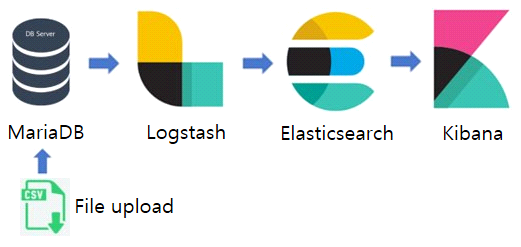
장채연 : 데이터 수집 / 데이터 정제 / ELK연동 / Kibana 시각화 / 보고서 작성

1. **Project 주제 : ELK를 활용한 실시간 공공데이터 분석**

한양대학교 데이터 분석 과정에서 교육받은 ELK를 활용하여 공공데이터를 실시간으로 분석해보는 프로젝트를 진행했습니다. 그중 해외에서 많이 이용하는 Bike Sharing이 각종 환경 등에 얼마나 영향을 받는지 궁금해져서 ELK에서 지원하는 기능들을 활용하여 보기 쉽게 Bike Sharing에 대한 이용 현황을 파악해보았습니다.

1. **개발 내용**

* **논리구성**



- ELK Stack 활용 : Logstash를 통해 CSV 파일을 업로드하여 Elasticsearch를 거쳐 Kibana에서 분석 및 시각화

- 데이터의 실시간 업로드를 위해 MariaDB를 연동

* **사용기술**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **이름** | **버전** | **설명** |
|  | **Elasticsearch** | 6.4.2 | -분산 검색엔진  -Logstash를 통해 수신된 데이터를  분석 및 저장 |
| **Logstash** | 6.4.2 | -오픈소스 서버측 데이터 처리  파이프라인  -데이터 수집 |
| **Kibana** | 6.4.2 | -데이터 시각화 및 실시간 분석 |
|  | **MariaDB** | 10.3 | 데이터 저장 |
|  | **Open JDK** | 1.8.0 | ELK 사용을 위한 JAVA 환경 |
|  | **WinSCP** | 5.11.3 | 파일 업로드 |
|  | **CentOS** | 7.6 | 서버 OS 환경 |

* **산출물**

1. Elastic(ELK) Stack 구축하기

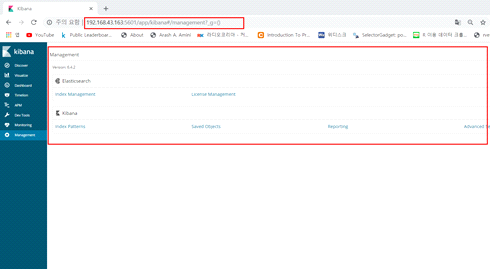


* ELK 란?

- ELK( ELK Stack )는 Elastic사 에서 제공하는 Elasticsearch, Logstash, Kibana의 세 가지 오픈 소스 프로젝트의 약자로, 각 제품이 연동되어 사용자가 서버로부터 원하는 데이터를 가져와 실시간 데이터에 대한 검색, 수집 및 분석 툴로서 동작.

- Elastic이라는 기업명에 걸맞게 높은 확장성과 뛰어난 이식성을 가지고 있어 다른 여러가지 툴 과도 연동이 가능.

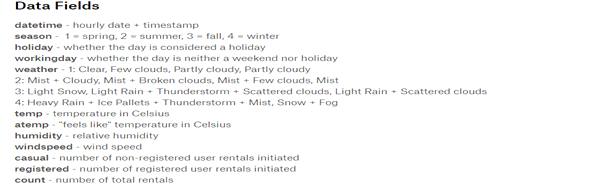
Elasticsearch와 Kibana 설치 이후 연결 확인



* About Data (Bike Sharing)
* 데이터 출처 -> 캐글 Bike Sharing Demand



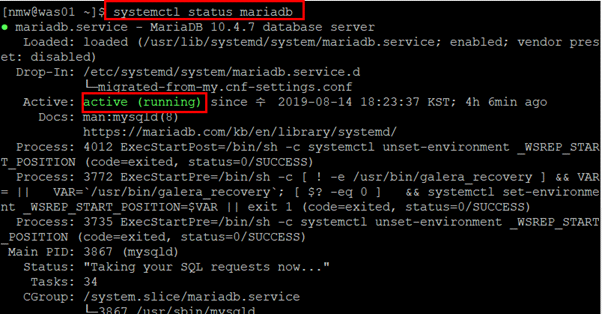
* 데이터 구조



1. DATA IMPORT TO MARIADB -> LOGSTASH(실시간)

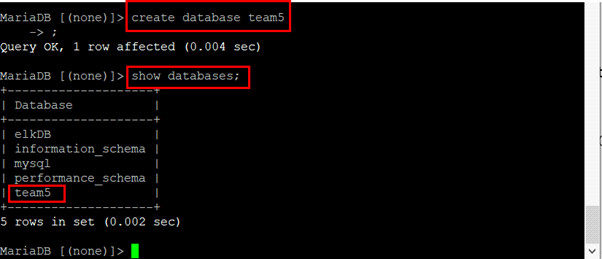
MariaDB설치 이후 및 연동

* MariaDB 시스템 작동확인

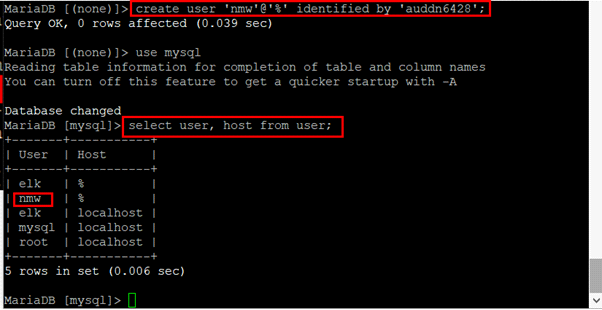


Logstash-Mariadb 연동

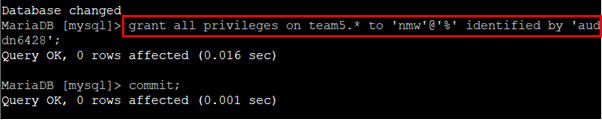
* Database team5 생성



User 생성 -> nmw



User 권한부여



User 권한확인

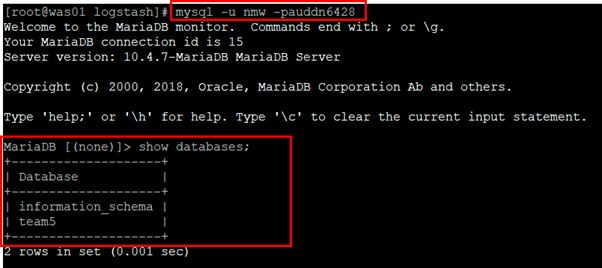


Table 생성

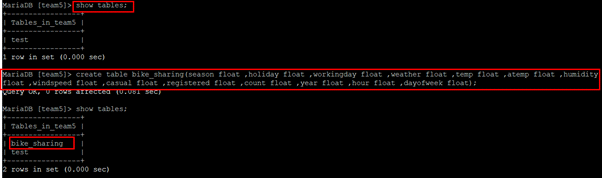
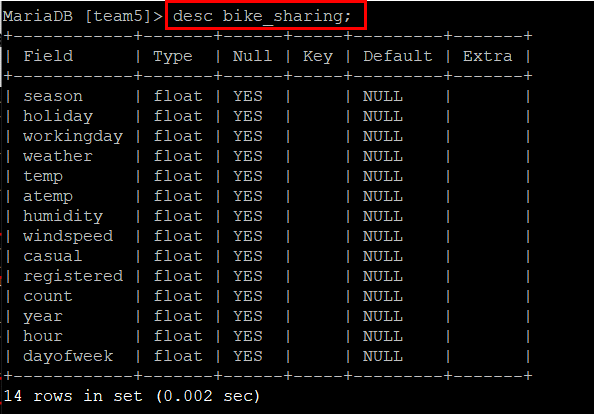
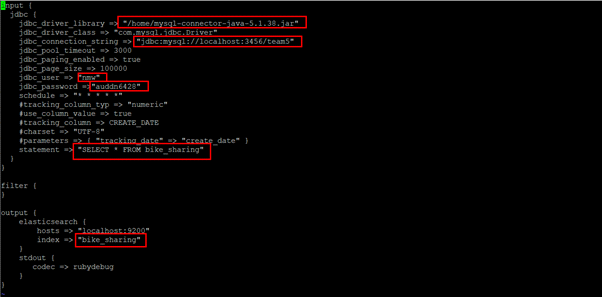


Table 구조확인



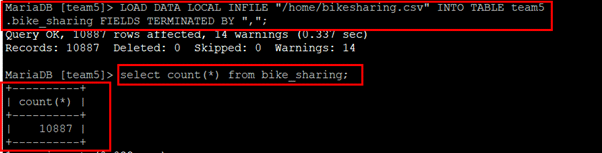
Logstash MariaDB\_config



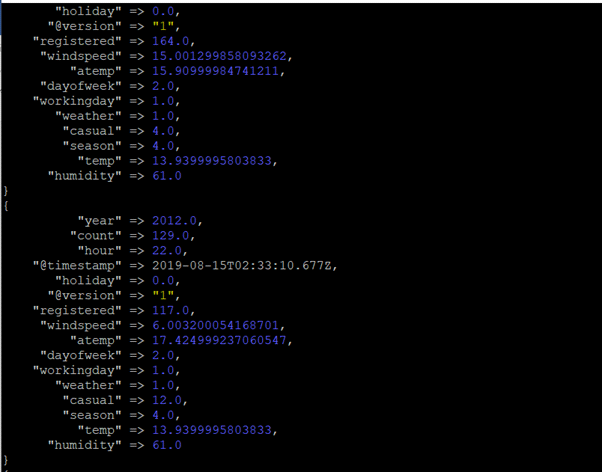
Logstash 실행



MariaDB data import

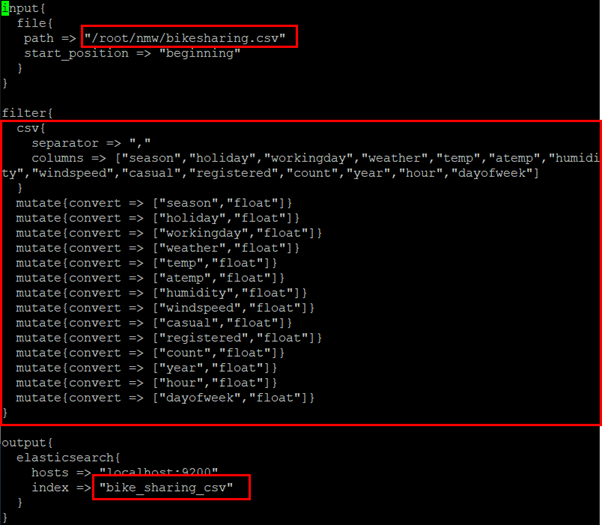


실시간 data import logging

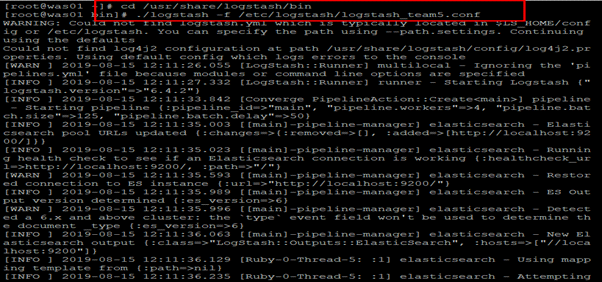


1. DATA IMPORT TO LOGSTASH DIRECTLY

Logstash config



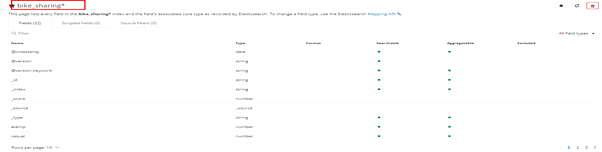
Logstash 실행



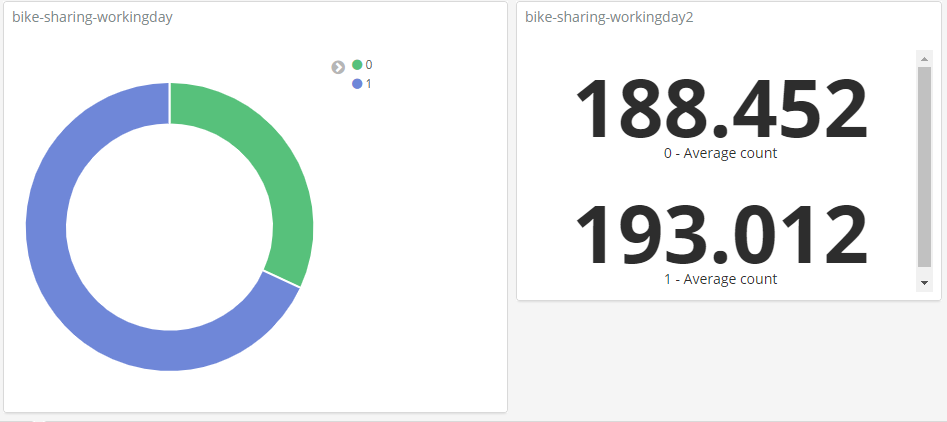
(elasticserch - index management)

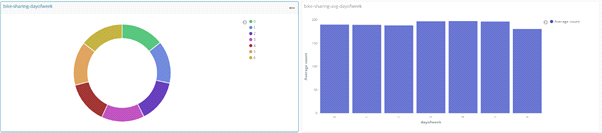
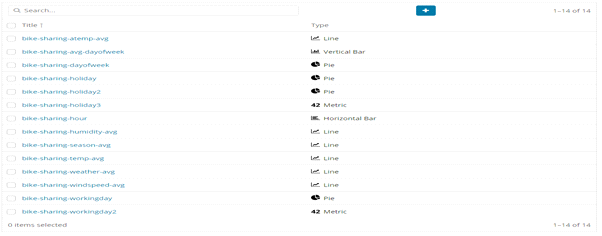
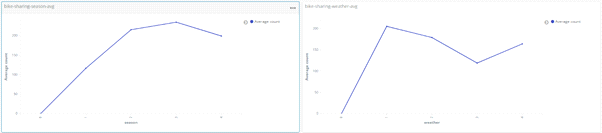
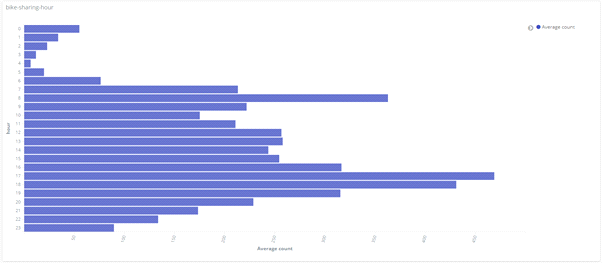


(kibana - index pattern)



1. Kibana를 활용한 <Bike-sharing Demand> data visualization 및 분석



* **프로젝트를 진행하면서…**
* 프로젝트를 진행하던 중, 추가 변동사항이 있을 때 서비스가 중단된다는 한계를 느껴서 data를 직접적으로 logstash에 업로드하는 방법에서 MariaDB를 통해서 logstash에 실시간으로 업로드하는 방법으로 변경해보았습니다.
* 이를 위해 Mariadb를 추가 연동하여 logstash에 data를 실시간으로 업로드할 수 있도록 하는 방법을 선택했는데, MariaDB와 logstash를 연동하는 과정에서 많은 어려움이 있었습니다.
* MariaDB를 설치하는 것부터 user 생성 권한관리, 각종 쿼리수행부터 logstash 와 연동하는 것까지 많은 시간을 소요했지만 이 과정을 통해 실시간으로 서비스 중단없이 데이터를 자유자재로 제어할 수 있다는 것을 확인하였습니다.
* 처음에는 Window OS에서 작업한 파일이나 다운로드 한 파일을 Virtualbox내의 CentOS에 공유하는 과정이 어려웠지만 반복되는 작업으로 WinSCP 라는 툴을 발견하였고, 그로 인해 어려움을 해결할 수 있었습니다.
* 다행히 데이터가 영어로만 이루어져 있어서 한글이 깨지는 오류를 피할 수 있었지만 앞으로 한글로 작성되어 있는 데이터를 사용하게 되어 오류가 생겼을 때 해결할 수 있도록 공부해야 할 것입니다.