육체 노동과 정신 노동에서 벗어나자!

빅데이터와 머신러닝을 통한 해킹과 보안

SangKeun Jang(maxoverpro)
http://www.maxoverpro.org
maxoverpro@gmail.com

우리에게 주어진 상황은?

IoT 기기들까지 … 보안을 해야 할 대상들이 점점 많아지고 있다. ㅜㅜ

고작해야 x명의 보안 전문 인력으로 보안을 다 하라고?!

이것저것 보안 솔루션들만 잔뜩 많아졌는데…잘 하고 있는 건가?

방화벽 로그

서버 로그

IPS/IDS 로그

네트워크 장비 로그

악성코드 로그

24시간이 모자라...

계획적으로 하는 일

정보보호 업무
(지침, 정책, ...)

정보 보안 교육

개인정보보호

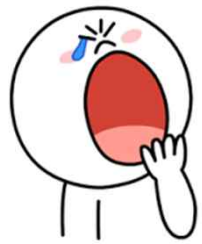
사이버테러대응 훈련

매일 하는 일

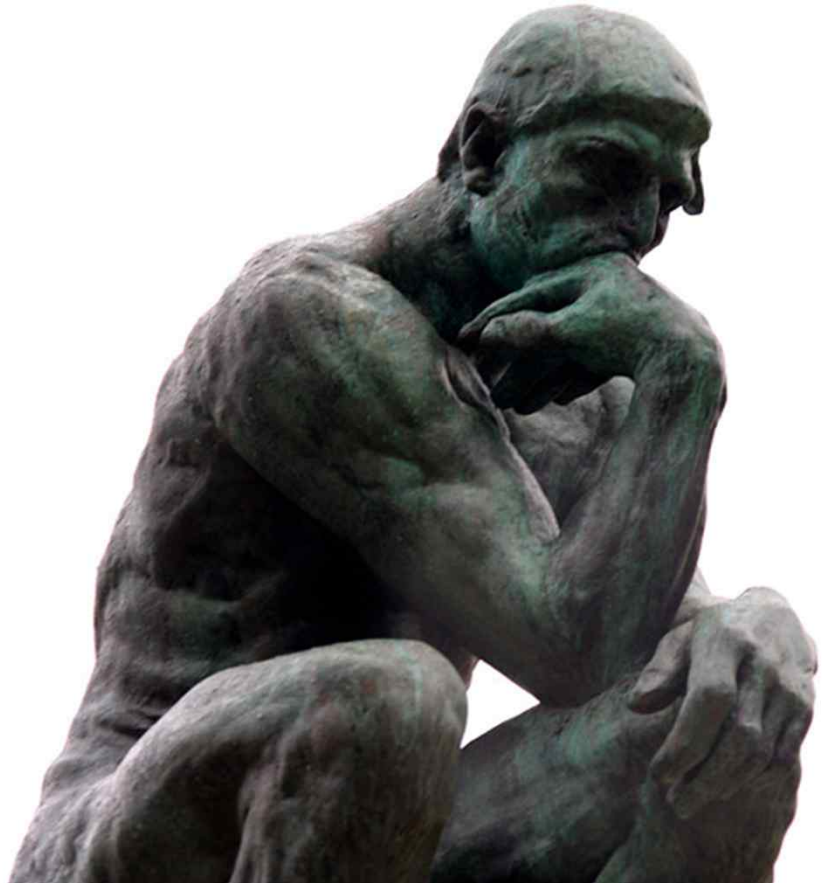
- ✓ 사고 예방
- ✓ 보안점검
- ✓ 위협 탐지, 분석 및 대응
- ✓ 보안 장비 운용
(Firewall, IDS/IPS, WAF, VPN ...)

반복적으로 매일 하는 일이란?

지루함은 특별히 할 일이 없을 때 개인이 느끼는 감정 상태를 뜻한다.
실행중인 일에 대해서 관심을 잃고 질려있는 것, 그리고 그 감정을 의미하기도 한다.
지루할 땐 졸림을 유발하는 경우가 많다.



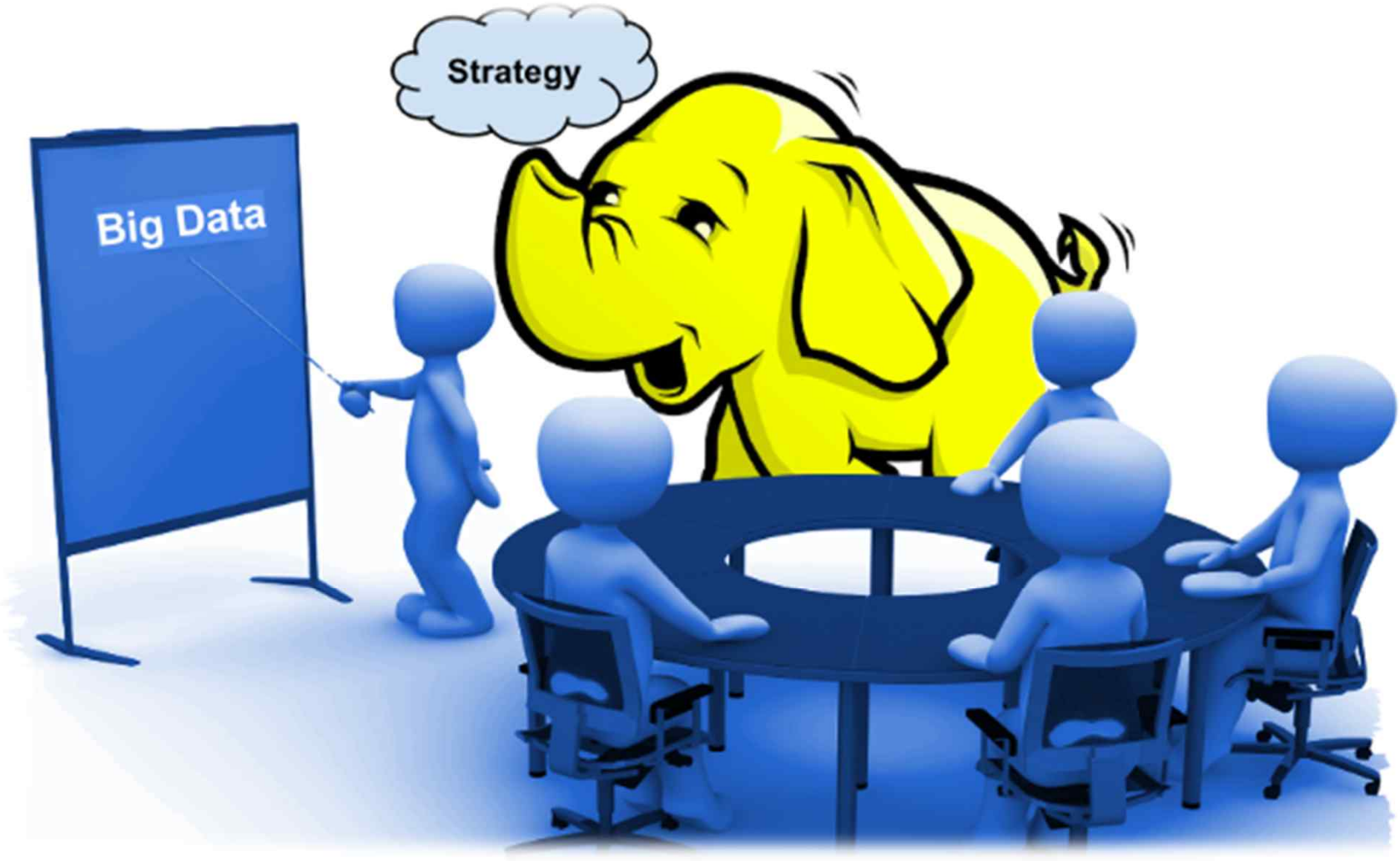
어떻게 해야 효율적으로
보안을 할 수 있을까?



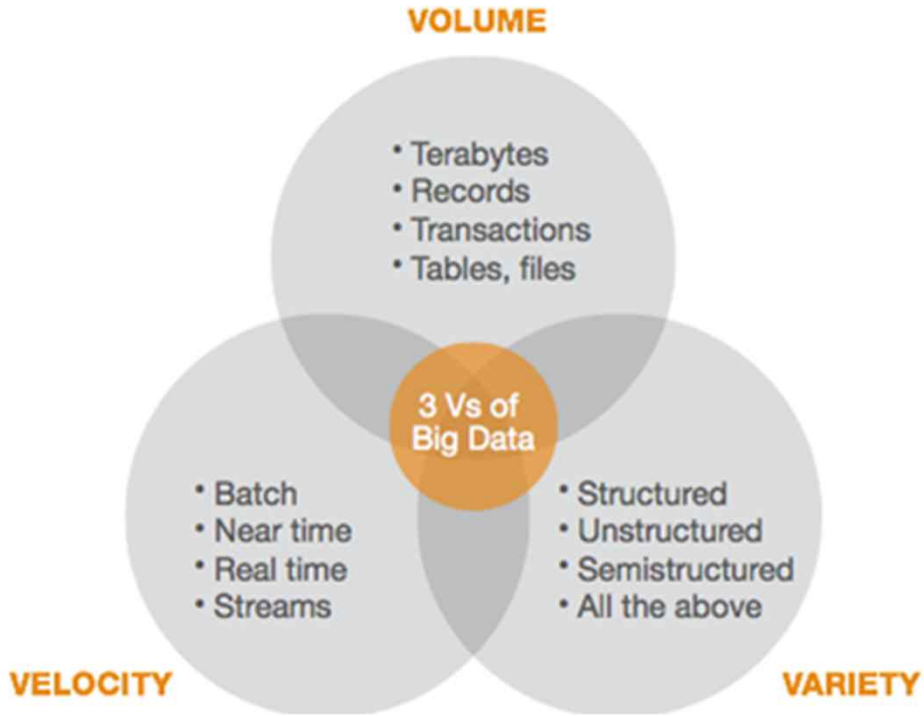
생각

1. 대용량 보안 로그 빠르게 검색
2. 자동화 된 보안 위협 탐지

어서와~코끼리는 처음이지?



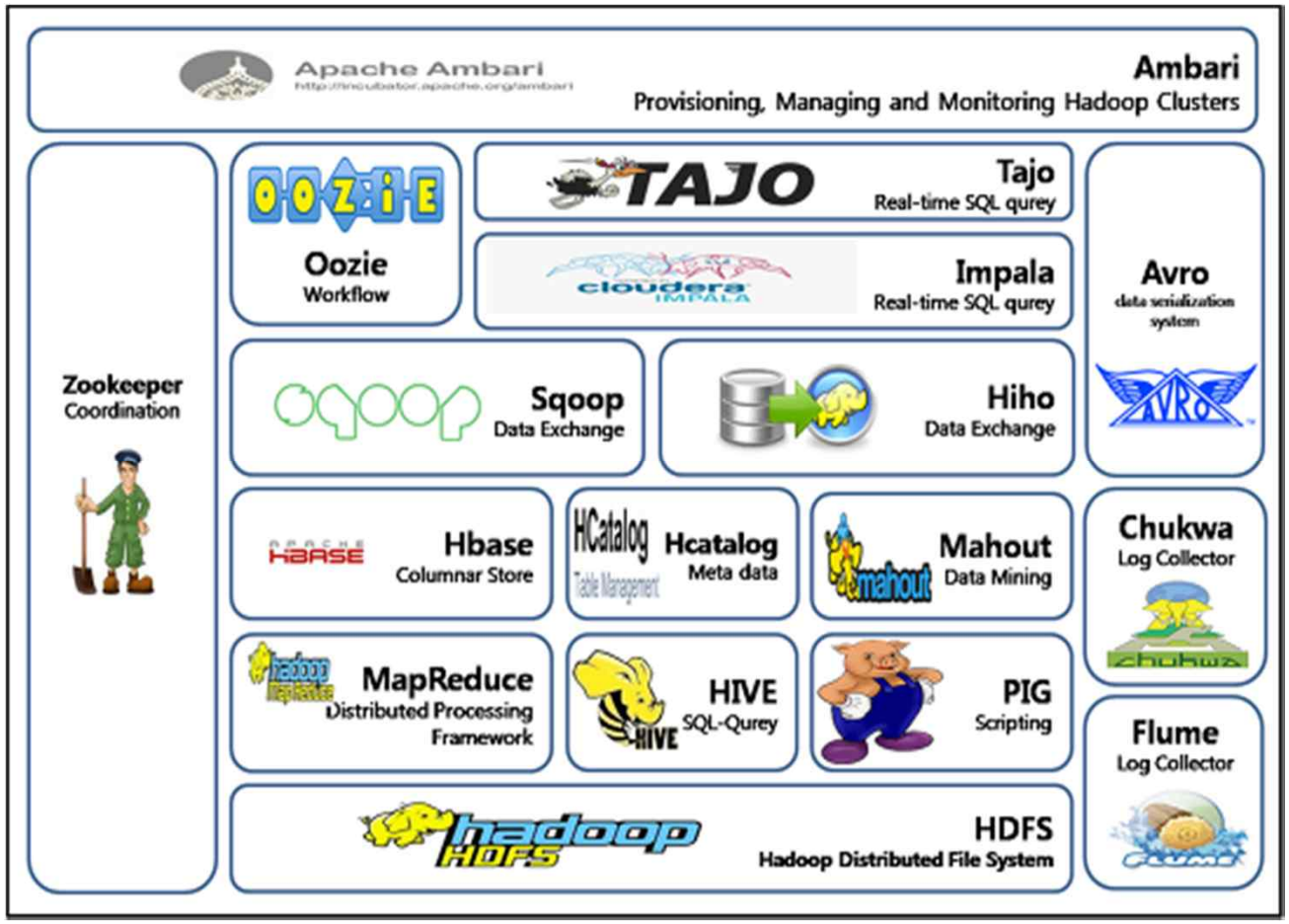
코끼리의 능력



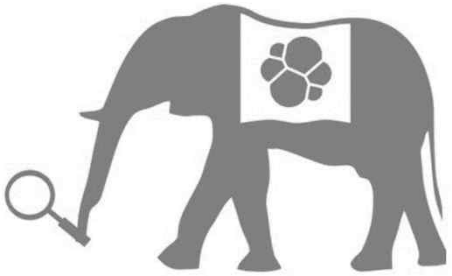
대용량(Volume)이고, 입출력도 빨라야 하고(Velocity), 다양한 데이터(Variety)를 처리해야 한다.

코끼리와 친구들

- 코끼리를 키우기 위해 동물농장을 운영할 수 없다면...



Apache Hadoop Ecosyster



Elasticsearch

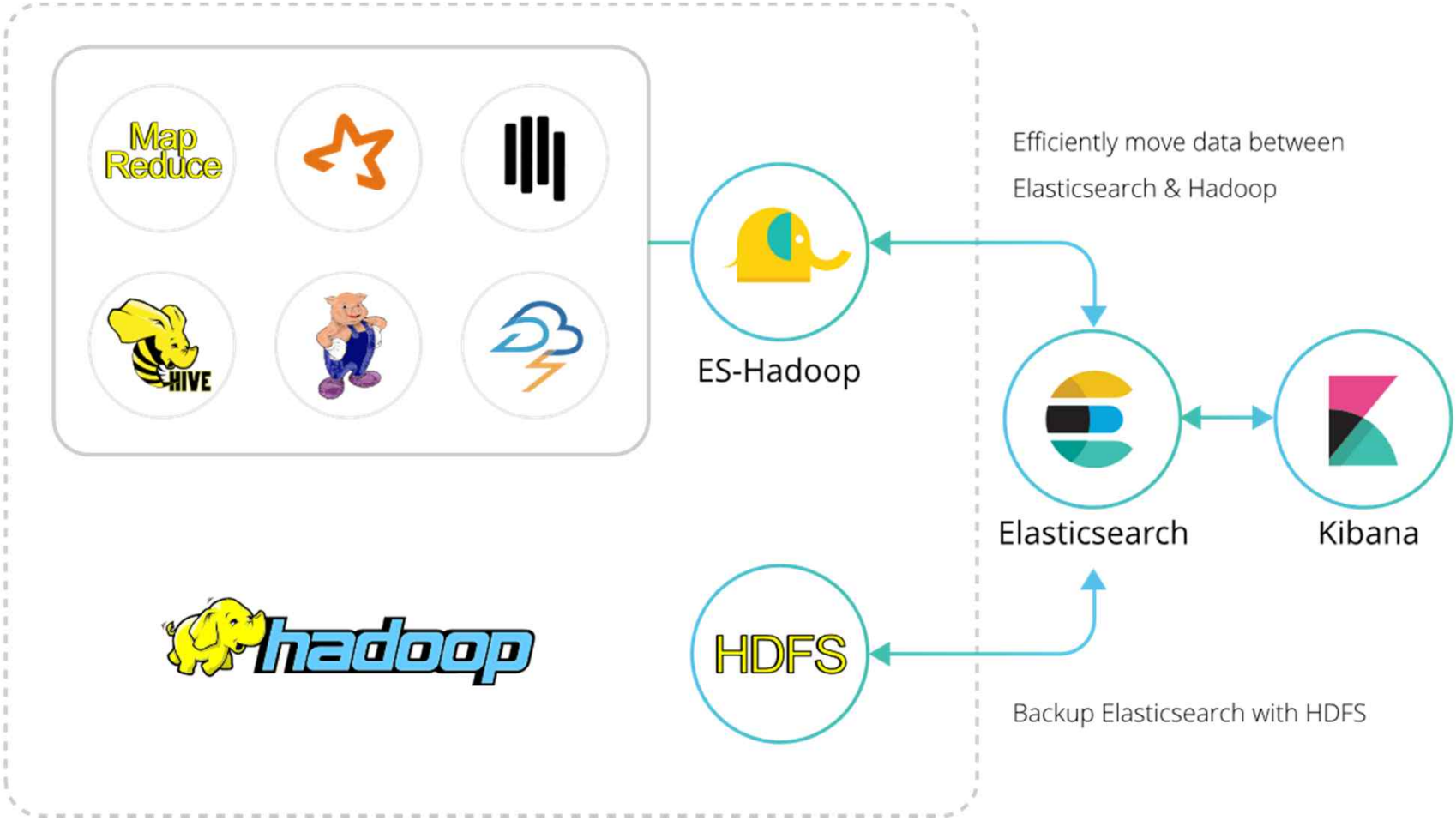
- Open Source.
- 데이터 저장소(PB급 데이터 저장 가능, 데이터 구조화)
- 내부 검색 엔진(Lucene)
- 실시간 분석 지원

- DB 와 뭐가 달라?

RDBMS(mysql…) -> elasticsearch
Database -> index
Table -> type
Row -> document
Column -> field

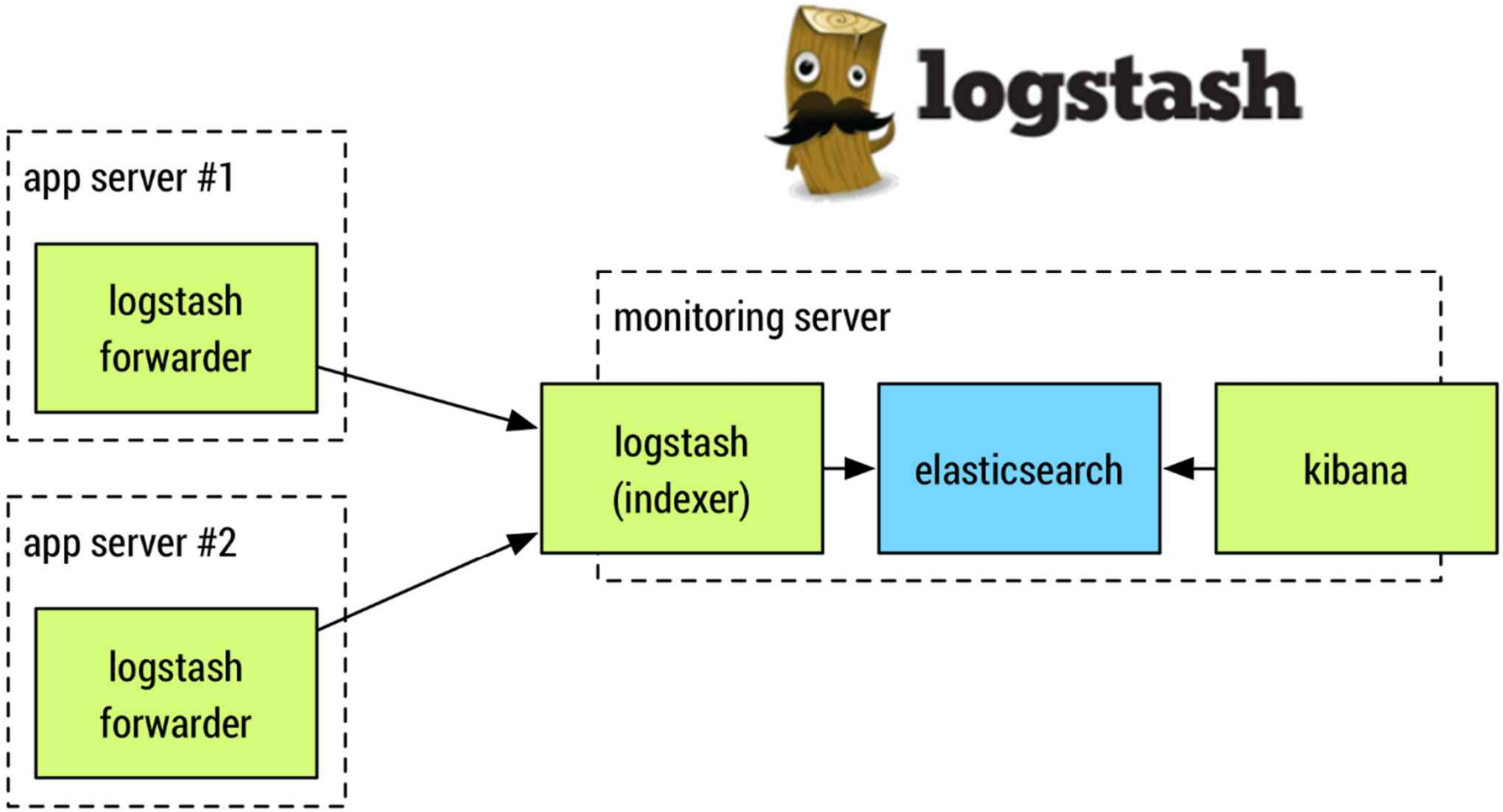
Elasticsearch 지원

- Elasticsearch 는 기존 Hadoop 과 연동이 가능.



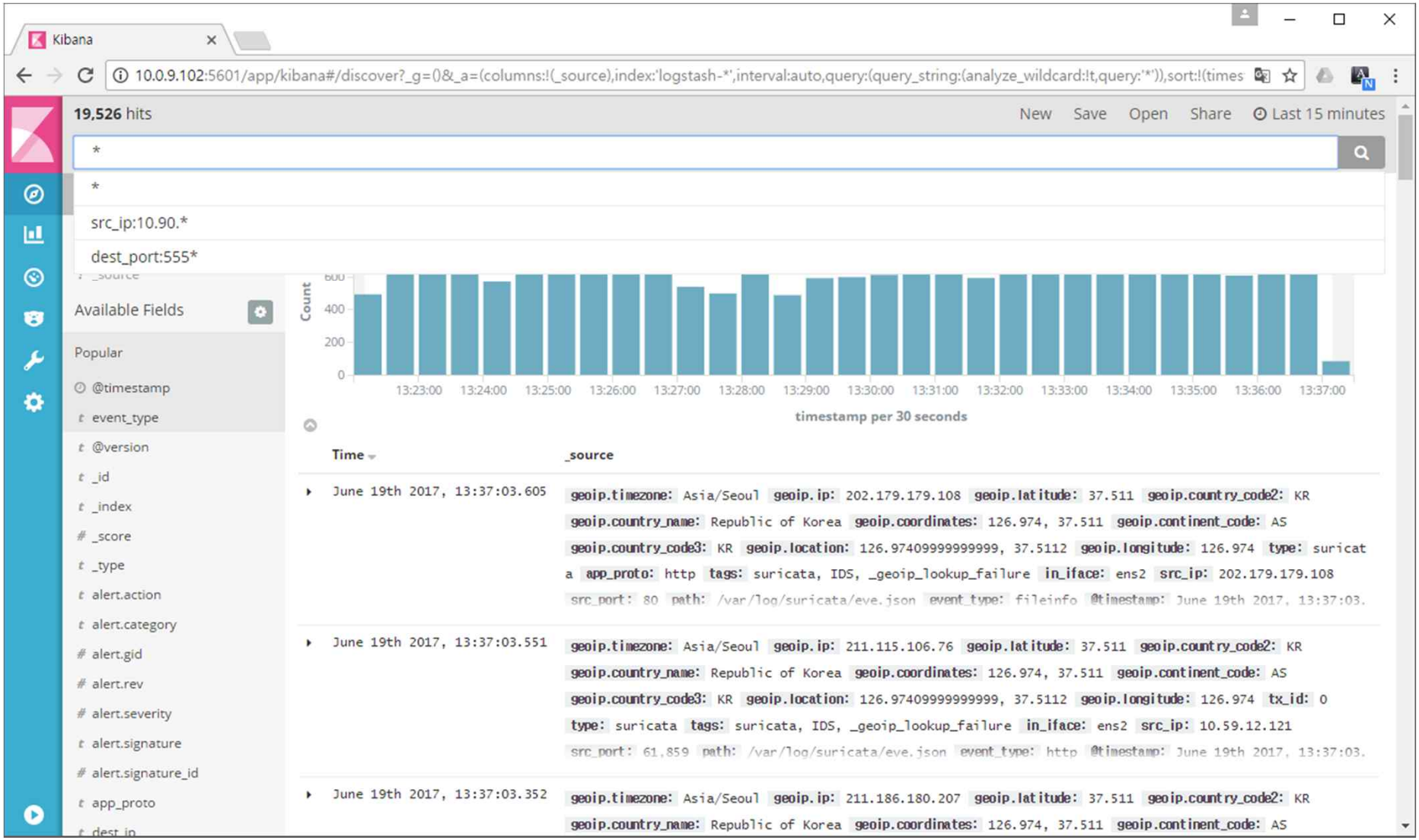
Logstash

- 각종 로그 수집을 위한 도구



Kibana

- 데이터 분석을 위한 시각화 된 도구.



코끼리 훈련 시키기

- 빅데이터의 처리 과정

문제 정의

데이터	내부, 외부, 미디어
수집	수동, 수집기, 크롤링
저장	정형, 비정형, 스토리지
처리	배치, 실시간, 스트리밍
분석	전처리, 통계, 마이닝
시각화	그래프, 차트, 테이블, 대쉬보드

코끼리 훈련 도구

- **빅데이터와 데이터 분석을 위한 도구(in python)**

- Python
- Numpy
- Pandas
- Matplotlib

- Scikit-learn
- Tensorflow
- Caffe
- CNTK
- Keras

데이터 전처리

- **보안과 관련된 데이터?**

  - Network
    Protocol, Header, IP, PORT, URL, Domain, Payload, …

  - File
    Signature,String, Contextual features, Instruction, …

  - Information privacy
    주민등록번호, 주소, 전화번호, E-mail…

데이터 분석

- 통계적 분석
  - 기술 통계량(산술 평균, 표준 편차, 최대값.최소값, 분산 등을 통해 산포도 확인)
  - 상관 분석(두 변수간의 관계 분석)
  - 회귀 분석(독립 변수와 종속 변수 사이의 상관 관계 분석)

- 데이터 마이닝
  - 데이터간의 상호 관련성 및 중요 정보 추출
  - 패턴인식, 인공지능, 고급 통계 분석

- 텍스트 마이닝
  - 텍스트 기반 데이터에서 정보 검색, 추출, 체계화, 분석 등의 기술을 통해 의미 있는 정보 추출

- 군집 분석
  - 비슷한 특성을 갖은 것들끼리 유사성을 발견하여 군으로 처리
  - 패턴인식, 인공지능, 고급 통계 분석

데이터 분석

- 3일 동안 악성코드 감염된 단말기는?

USER_A 가 3일 연속 악성코드 감염 현상 발생됨.

	date	node
0	20170405	USER_A
1	20170405	USER_B
2	20170405	USER_C
3	20170406	USER_A
4	20170406	USER_A
5	20170406	USER_B
6	20170406	USER_D
7	20170406	USER_E
8	20170406	USER_E
9	20170406	USER_F
10	20170406	USER_G
11	20170407	USER_A
12	20170407	USER_H
13	20170407	USER_I
14	20170407	USER_H

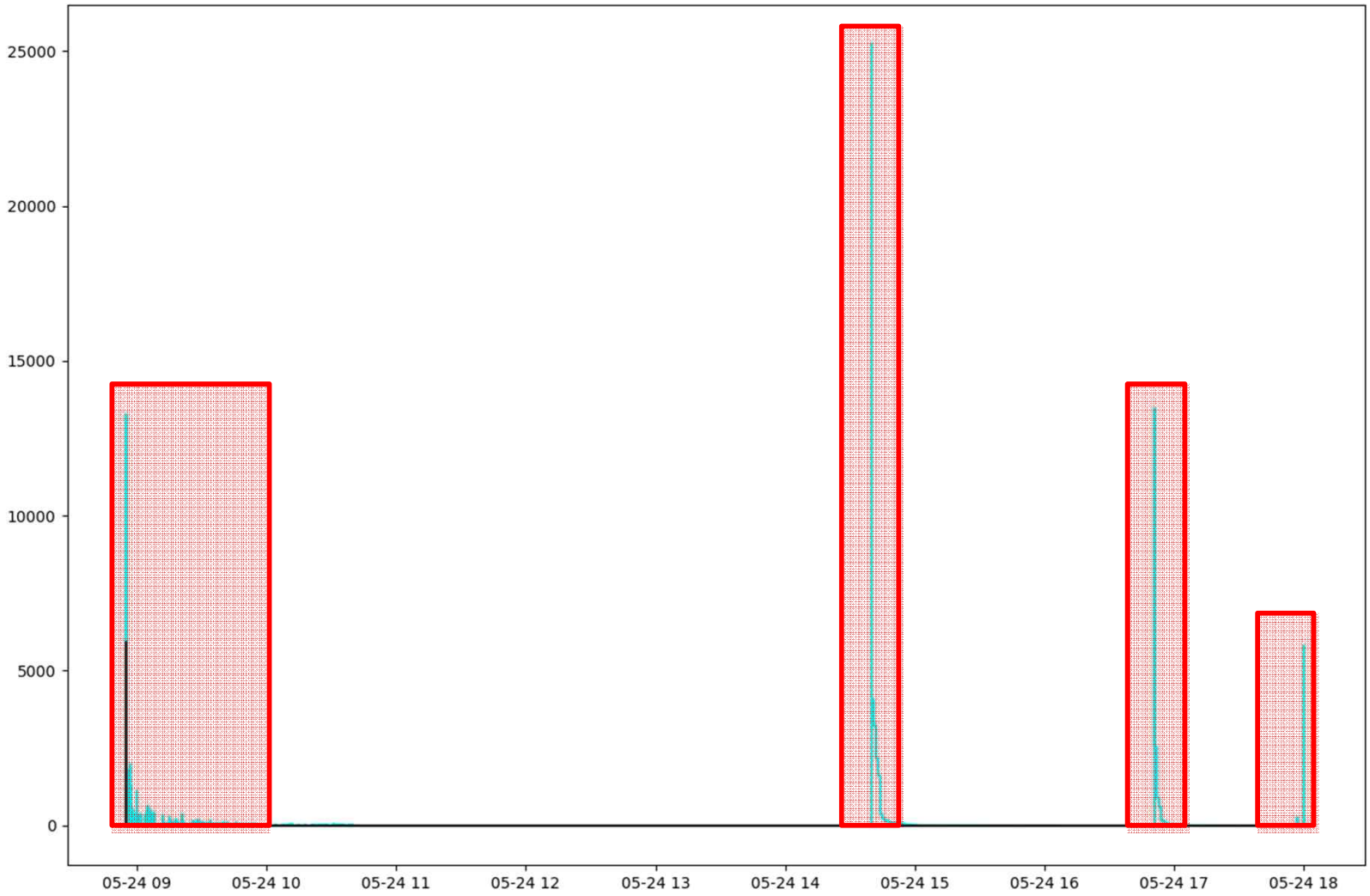
date	20170405	20170406	20170407	All
node				
USER_A	1	1	1	3
USER_B	1	1	0	2
USER_C	1	0	0	1
USER_D	0	1	0	1
USER_E	0	1	0	1
USER_F	0	1	0	1
USER_G	0	1	0	1
USER_H	0	0	1	1
USER_I	0	0	1	1
All	3	6	3	12

=== 3일 연속 악성코드 감염자 ===

date	20170405	20170406	20170407	All
node				
USER_A	1	1	1	3

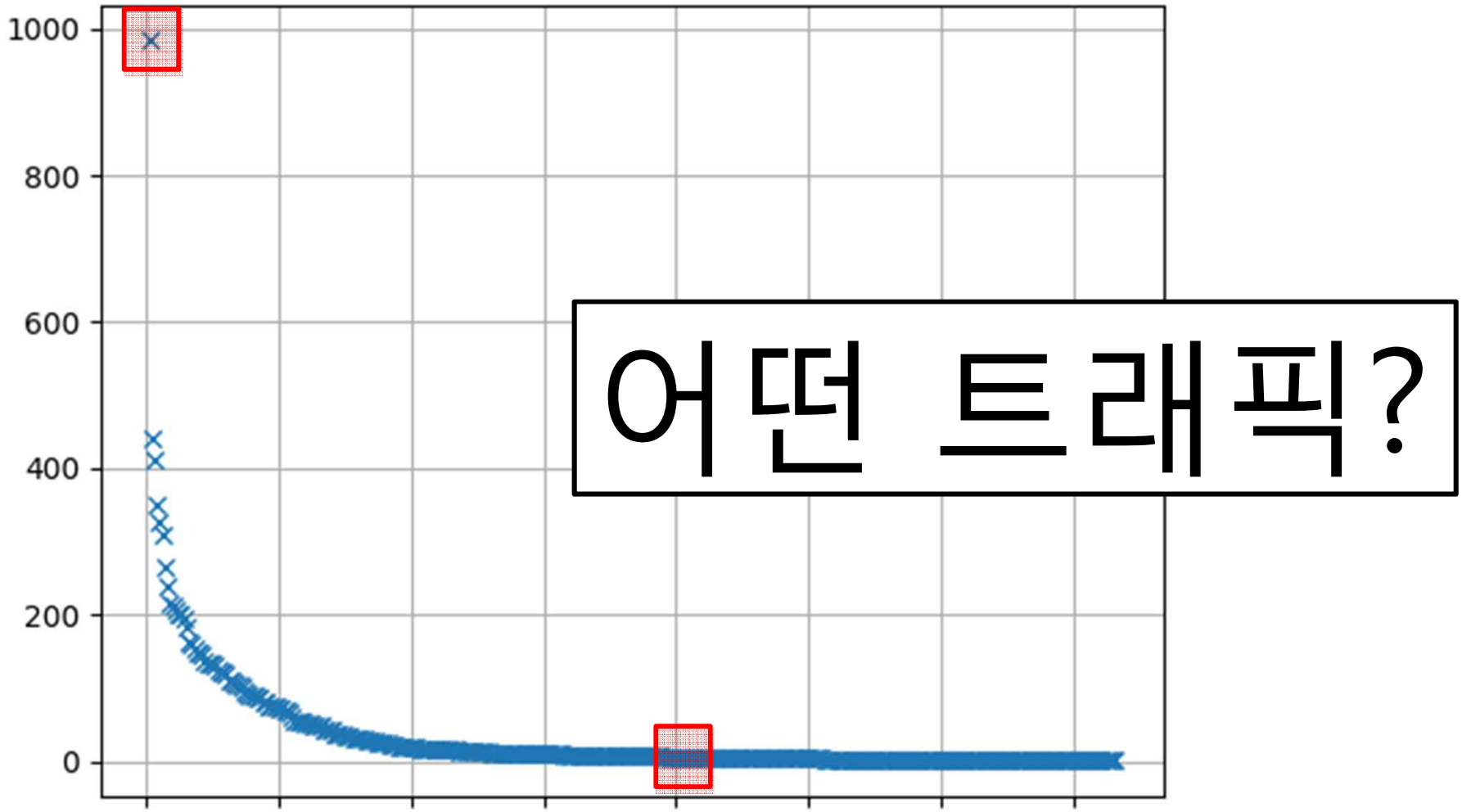
데이터 분석

- USER A의 단말기는 언제부터 얼마나 네트워크를 사용하는가?



데이터 분석

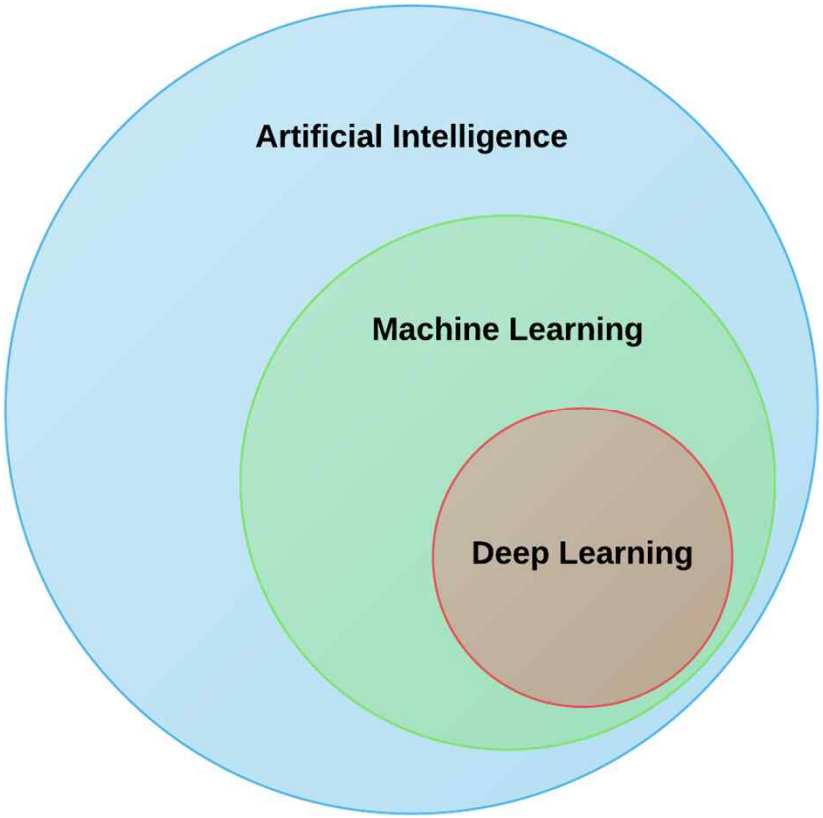
- USER A 의 트래픽 형태



인공지능(AI)?

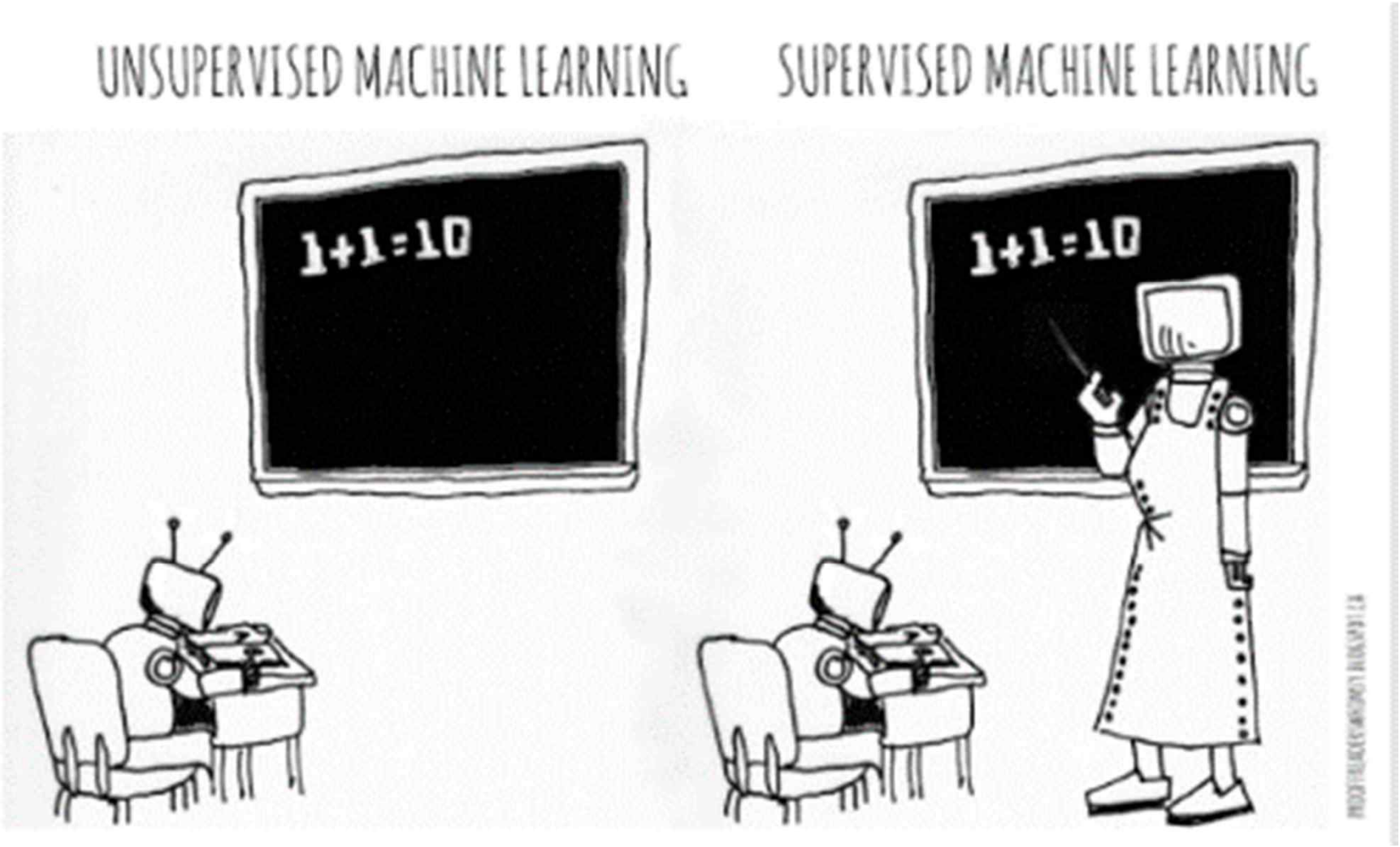
- **AI? ML? DL? 의 차이점이 뭘까?**

  - AI : 폭 넓은 의미
  - ML(Machine Learning) : 프로그램 된 데이터를 스스로 학습하여 분석&예측.
  - DL(Deep Learning) : Deep Neural Network 를 통해 학습



머신러닝

- 머신러닝 종류


UNSUPERVISED MACHINE LEARNING
1+1=10
SUPERVISED MACHINE LEARNING
1+1=10

머신러닝

- 머신러닝 종류


지도 학습
(Supervised Learning)
비지도 학습
(Unsupervised Learning)
강화 학습
(Reinforcement Learning)
분류
(Classification)
예측
(Prediction)
군집
(Clustering)
State
Action
Reward
kNN
SVM
Decision Tree
Naïve bayes
...
Linear
regression
K-means
DBSCAN
Brute force
Monte carlo
MDP
Q-Learning
...

생각해봐야 할 것들.

- 이미 공격자가 침입한 상황이라면?
- 데이터가 많거나 적거나 한 상황에서는?
- 데이터가 편향 되거나 변화가 많은 경우는?
- 공격자가 AI 기술 사용한다면?
- 정상 파일의 난독화 된다면?