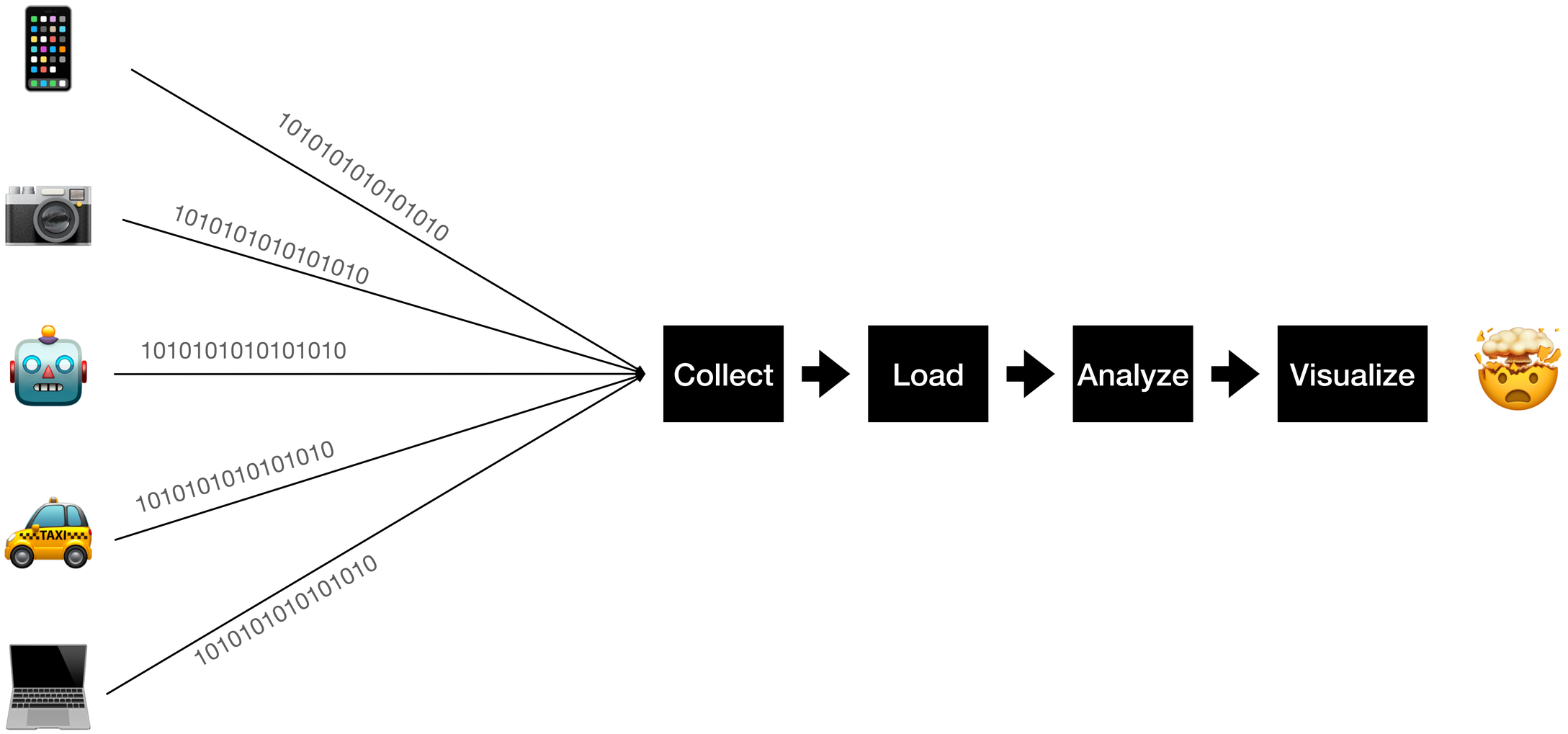
Big Data Analytics Programming

Week-13. Elastic Stack

Jungwon Seo, 2020-Fall

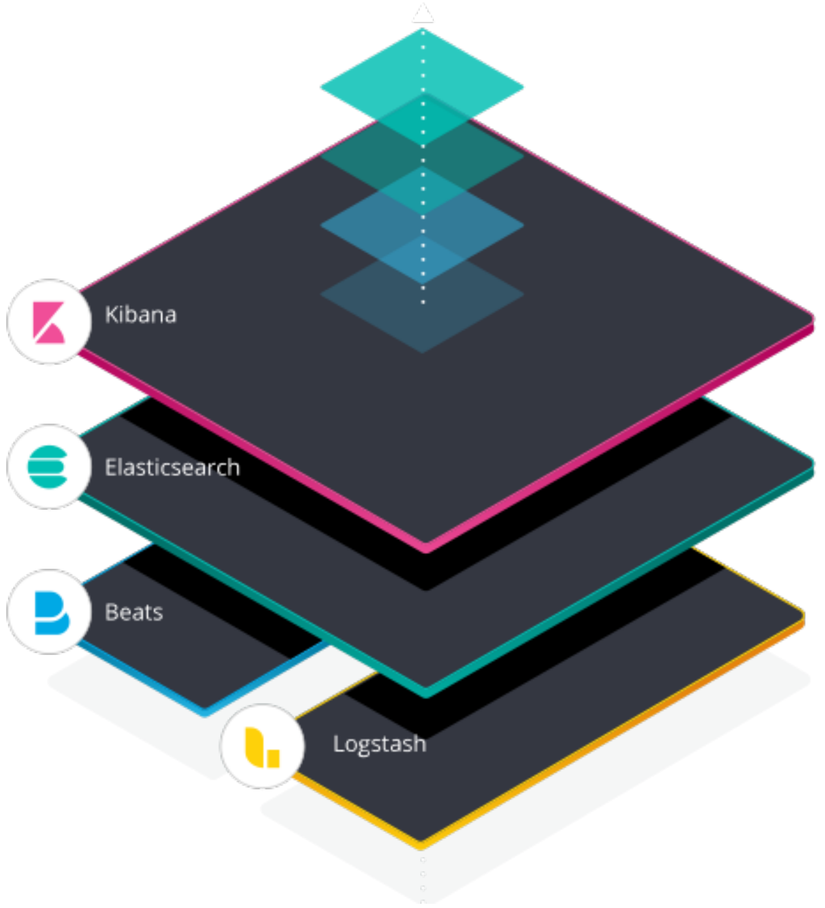
여러 곳에서
동시 다발적으로
대량 생산되는 데이터를
어떻게
수집-저장-분석-시각화 할 수 있을까?


1010101010101010
1010101010101010
1010101010101010
1010101010101010
1010101010101010
Collect
Load
Analyze
Visualize

ELK Stack

Elasticsearch, Logstash, Kibana

- Elasticsearch
  - 검색 및 분석 엔진
- Logstash
  - 데이터 처리 파이프라인
  - 여러 소스에서 동시에 데이터를 수집 및 변환
- Kibana
  - 차트와 그래프를 이용해 데이터 시각화
- Beat
  - 파일 추적 (Log 파일)
  - Beat가 추가됨으로써 ELK Stack에서 **Elastic Stack**으로!



https://www.elastic.co/kr/what-is/elk-stack

Elasticsearch

Elasticsearch

오픈소스 검색엔진

- Lucene 라이브러리 기반의 검색엔진
- REST API 형태로 접근
  - HTTP : PUT/DELETE/GET/POST
- 가장 대중적인 엔터프라이즈 검색엔진
- 최근 **ELK(Elasticsearch, Logstash, Kibana)** 스택이라는 빅데이터 수집 및 분석에 많이 사용됨

Elasticsearch

RDB vs ElasticSearch

Term	Document
Big	Doc1, Doc2, ..
Data	Doc1, Doc3, …
…	…

Elasticsearch: O(1)

Document_id	Content
Doc1	Big data is very big
Doc2	Data science is science
…	…

RDB: O(n)

Seach : "Big"

Elasticsearch

RDB vs ElasticSearch

관계형 데이터베이스 (mysql)	엘라스틱서치
Database	Index
Table	Type
Row	Document
Column	Field
Schema	Mapping
Index	모두 Index되어있음
SQL	Query DSL

Elasticsearch Mapping

Mapping이란?

- 관계형 데이터의 스키마와 동일
- Mapping 없이 데이터를 엘라스틱서치에 삽입?
  - 가능
  - 하지만, Mapping을 없이 데이터를 넣는 것은 데이터의 사용성이 떨어질 수 있음
    - 예를들어 2020-11-07이라는 값을 Mapping없이 넣는다면?
      - String으로 인지할 수 있음
      - 날짜별 정렬, 월별 정렬/필터링과 같은 연산을 사용할 수 없음
    - 숫자를 입력했지만, 문자열로 인식했다?
      - Min, max, mean, median과 같은 연산을 사용할 수 X
- 가능하다면! 항상 Mapping을 먼저 지정해 놓고 데이터를 삽입!
  - 데이터 먼저 넣고, Mapping을 후에 지정해줄 수도 있음

Elasticsearch Search

Search 방식

- request_body에 json 형식으로 조건을 작성!

```
body = {
    "query": {
        "term": {
            "points":30
        }
    }
}
res = es.search(body=body,index=INDEX_NAME)
pprint.pprint(res)
```

Elasticsearch Aggregation

Aggregation 이란?

- Search Query 사용시 수치적 값들의 다양한 값들을 얻어 낼 수 있다.

```
body = {
    "size" : 0,
    "aggs" : {
        "avg_score" : {
            "avg" : {
                "field" : "points"
            }
        }
    }
}
res = es.search(body=body,index=INDEX_NAME)
pprint.pprint(res)
```

Elasticsearch Bucket Aggregation

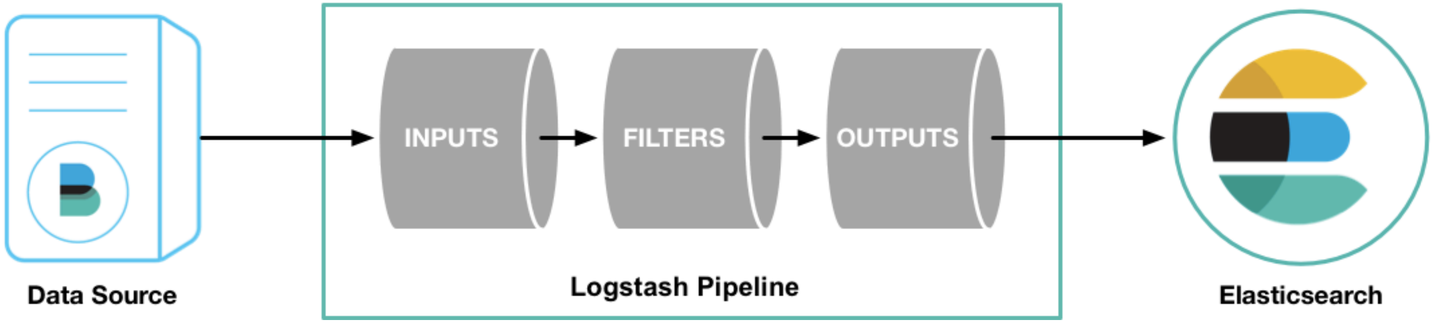
Bucket Aggregation 이란?

- RDB의 group by와 유사한 기능
- “document의 bucket을 만든다”

```
body = {
    "size" : 0,
    "aggs" : {
        "team_stats" : {
            "terms" : {
                "field" : "team"
            },
            "aggs" : {
                "stats_score" : {
                    "stats" : {
                        "field" : "points"
                    }
                }
            }
        }
    }
}
res = es.search(body=body,index=INDEX_NAME)
pprint.pprint(res)
```

Logstash

Input, filters, outputs



https://www.elastic.co/guide/en/logstash/current/first-event.html

E.O.D