

MAIN PROJECT





# OVERVIEW OUR TEAM TIMELINE RESULT REVIEW

배경

팀 구성

수행 절차

수행 결과

소감



# OVERVIEW

도메인

주제

목적

개요



# KOHUS

© CATENOID



# VOD 서비스 이용 정보 분석 시스템

## REQUEST

- ① 새로운 인사이트
- ② 고객친화
- ③ 상용화
- ④ 최소비용 최대효율

## RESTRICTION

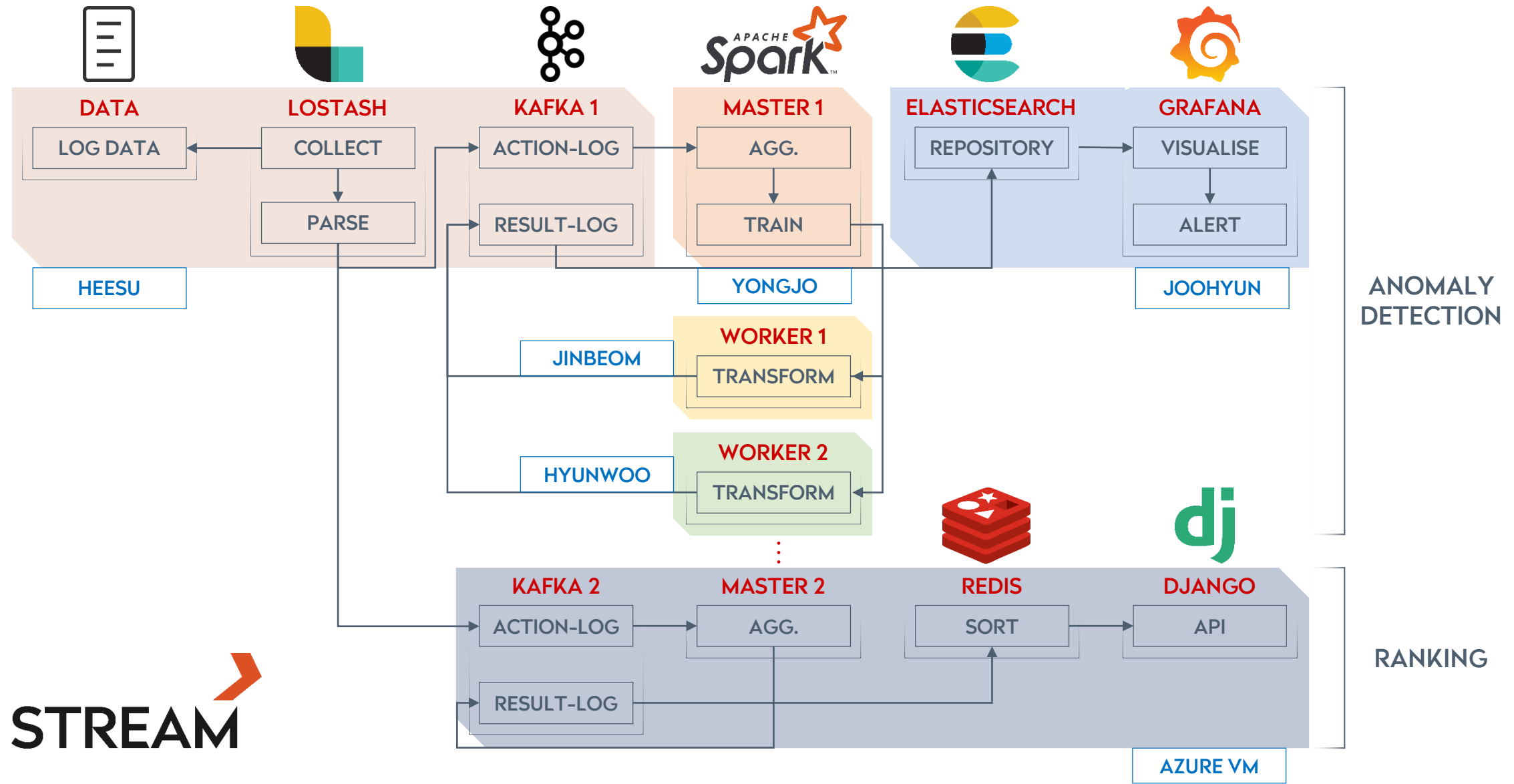
- ① 실시간
- ② 처리량 (일 30GB ↑)
- ③ 자동화
- ④ 오픈소스
- ⑤ 연관성 분석 지양

## WHY?

- ① 메타데이터 부재
- ② 고객 핵심자산 접근 불가
- ③ 기존 방식의 한계
- ④ '비정형', '빅' 데이터

# 대용량 데이터 실시간 분석 시스템

고객에 대한 지식이 없더라도,  
GB데이터가 밀리세컨드 단위로 들어오더라도,  
누구나 쉽고, 빠르게, 직관적으로 이해할 수 있는  
BASE 이상감지, 랭킹 파이프라인





# OUR TEAM

WADO  
R&R



DAO > DAO

다오. “당신들이 원하는 바를 다오(DAO). 우리가 줄게.”



# MEET OUR TEAM.



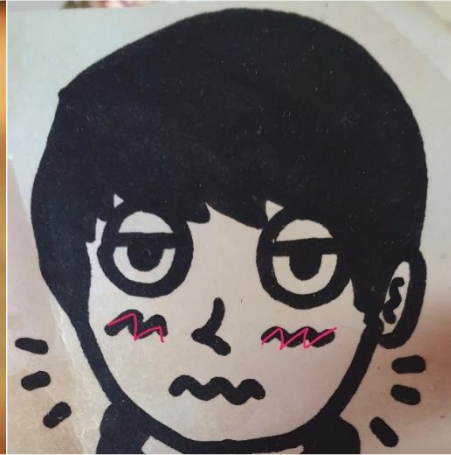
김주현

PM



신희수

PL



김용조

ENGINEER



김진범

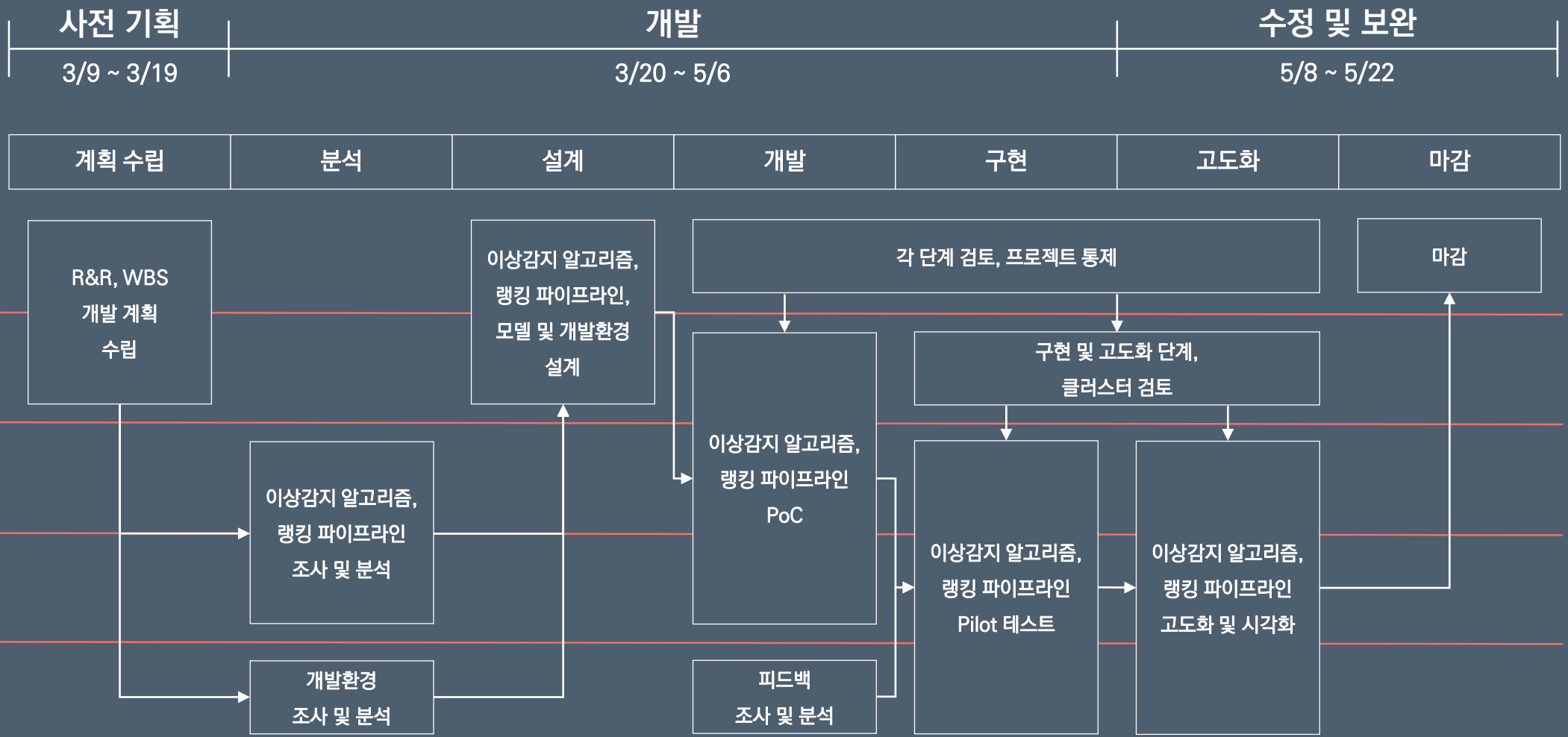
ANALYST



이현우

RESEARCHER



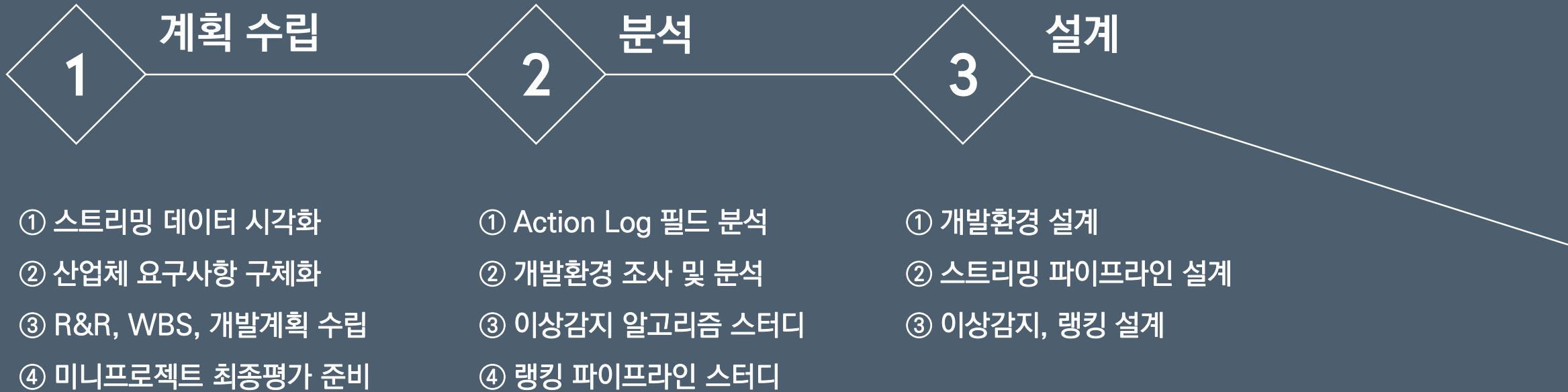




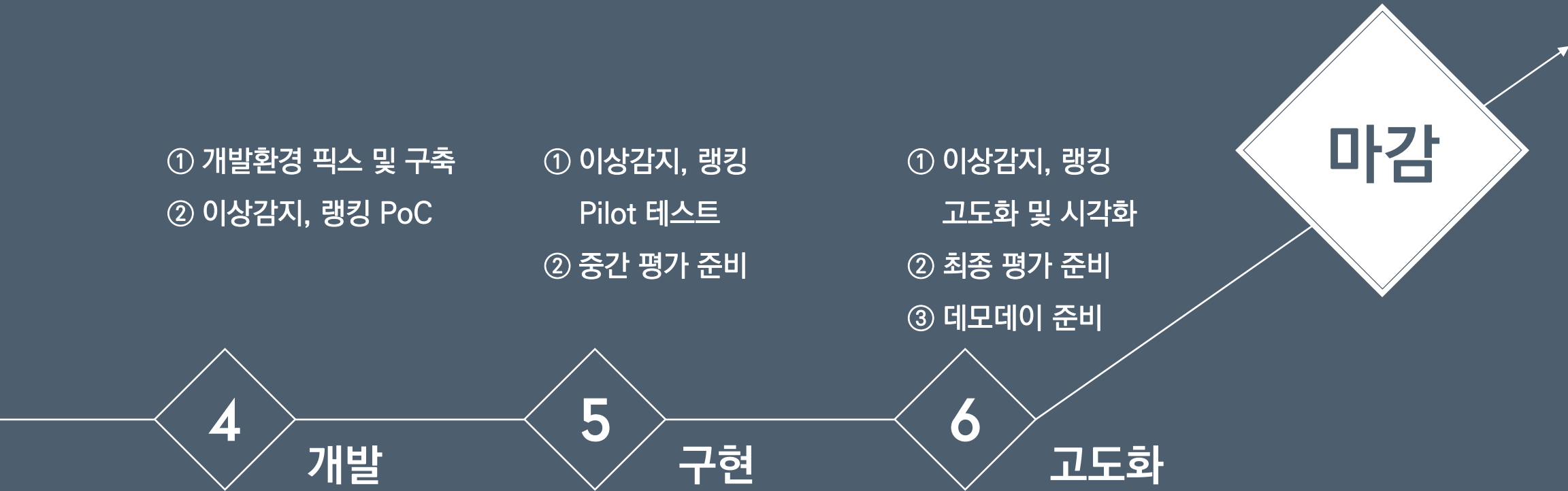
# TIMELINE

WORK FLOW











# RESULT

분석  
전처리  
이상감지  
랭킹



```
{ "_index": "action-2020.03.03", "_type": "useractionlog", "_id": "AXCdsUSHFx_XC
L_vJThy", "_score": null, "_source": { "@fields": { "action": "v
", "method": "vgrender", "ip_address": "211.252.200.1", "user
_agent_string": "Mozilla/5.0 (Windows NT 10.0; WOW64; AP
CPMS=^N20171204015852983469C0C765FEDFE3E23F_4266^; Tride
nt/7.0; rv:11.0) like Gecko", "d_bm1": 0.1263, "d_bm2": 0.0
006, "d_bm3": 0.0003, "d_bm4": 0.0319, "d_bm5": 0, "d_bm6": 0.01
14, "params": { "mid": 8750185, "mck": "5mocrazu", "base_mck": "
5mocrazu", "cid": "null", "category_id": 79271, "m": "kr", "msi
": 1, "cpk": "ttschool", "cpid": "1768", "mpk": "8hlgzg11", "hou
r": "0", "day": "3", "day_of_week": "2", "week": "10", "month": "
3", "year": "2020", "hit": 1, "pfkey": "ttschool-pc1-
hd", "islive": 0, "title": "180829_\uc57c\ub098\ub4500T_\uae
30\ucd08\ud68c\ud654STEP1", "domain": "www.yanadoo.co.kr",
"referer": "https://www.yanadoo.co.kr/common/playerSt
udy.do", "ofilename": "180829_\uc57c\ub098\ub4500T_\uae30\
ucd08\ud68c\ud654STEP1.mp4", "ufkey": "20181126-
1rnwbfo", "cname": "E02_1\uc5b4\uc21c\uac10\uac01", "uploa
d_time": 1543214230, "xid": "613a71daa8de445ef0be1a8aeaa8e0
5a3a047a622e612b90d9833e55bd075e34"}, "custom_log_keys": {
"custom_log_key0": "null", "custom_log_key1": "null", "custo
m_log_key2": "null", "custom_log_key3": "null", "custom_log_
key4": "null", "custom_log_key5": "null", "custom_log_key6":
"null", "custom_log_key7": "null", "custom_log_key8": "null"
, "custom_log_key9": "null"}}, "ip_address": "211.252.200.1"
, "user_agent_string": "Mozilla/5.0 (Windows NT 10.0; WOW
64; APCPMS=^N20171204015852983469C0C765FEDFE3E23F_4266^;
Trident/7.0; rv:11.0) like Gecko", "logdate": "2020-03-
03T00:00:00+00:00", "@version": "1", "@timestamp": "2020-03-
03T00:00:00.000Z", "path": "\\var\\log\\ual\\ual.log", "hos
t": "kr01dk16", "type": "useractionlog", "useragent": { "name"
: "IE", "os": "Windows 10", "os_name": "Windows 10", "device":
"Other", "major": "11", "minor": "0"}, "geoip": { "ip": "211.252
.200.1", "country_code2": "KR", "country_code3": "KOR", "coun
try_name": "Korea, Republic of", "continent_code": "AS", "la
titude": 37, "longitude": 127.5, "timezone": "Asia/Seoul", "l
ocation": [127.5, 37], "isp": "Korea Telecom-
PUBNET"}}, "sort": [1583193600000]} }
```

## 우리가 알고 싶은 것

- 어떤 무언가가 “hit”에 영향을 미치는가? **X**
- 고객들의 “hit”의 상태가 지금 이상한가? **O**
- 고객의 어떤 콘텐츠의 “hit”가 지금 핫한가? **O**

## 우리의 데이터

Kollus로 VOD 재생 시 발생하는 비정형 데이터

## 우리에게 필요한 데이터

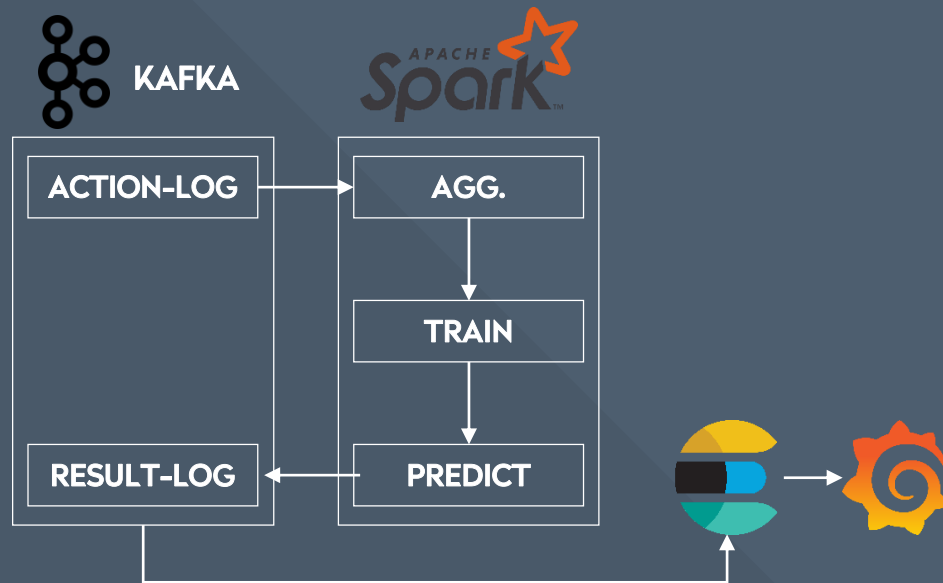
구분 가능한 키, 셀 수 있는 밸류가 있는 시계열 데이터

**LOGSTASH**  
**INPUT: FILE**  
**OUTPUT: KAFKA**

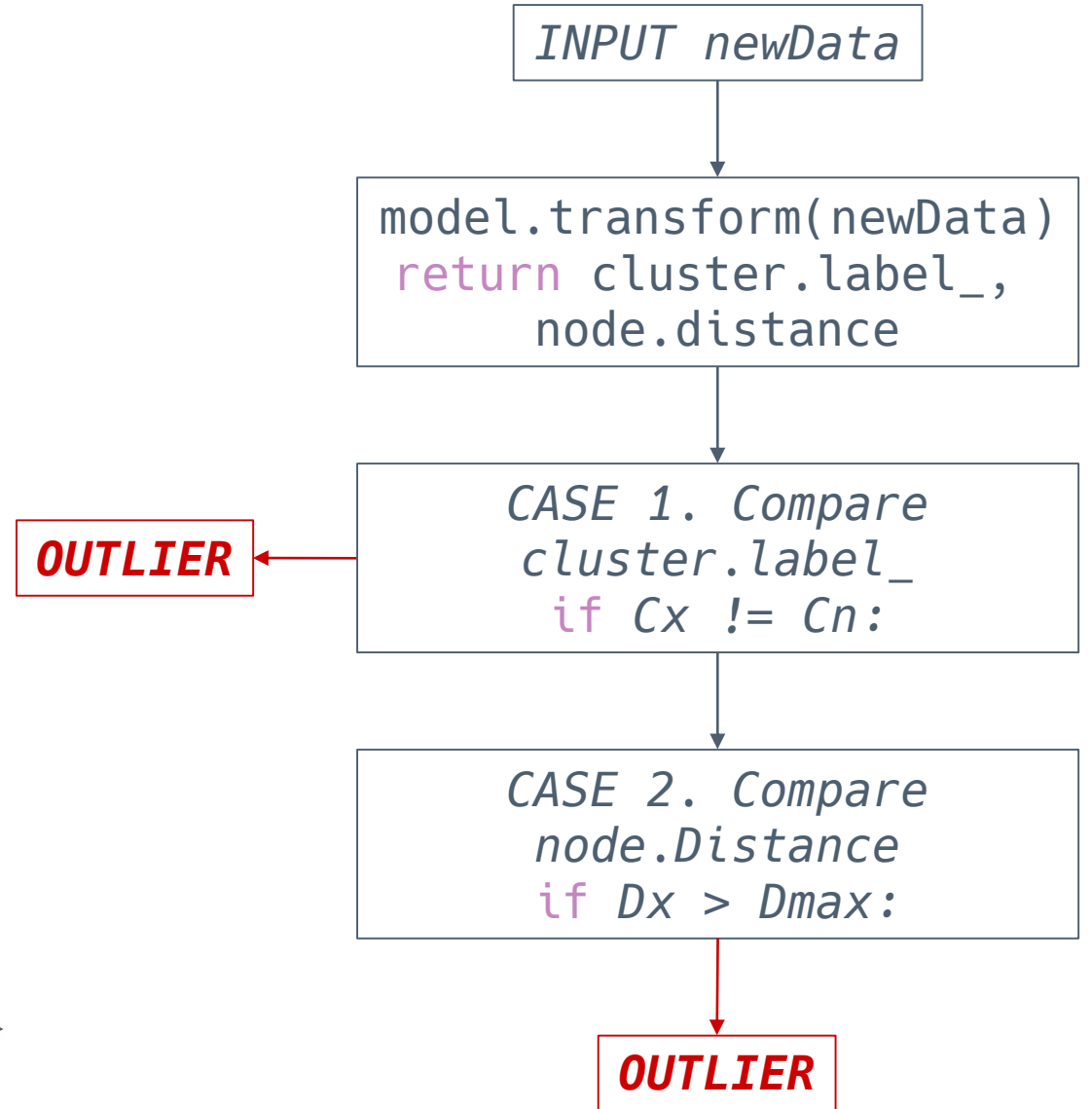
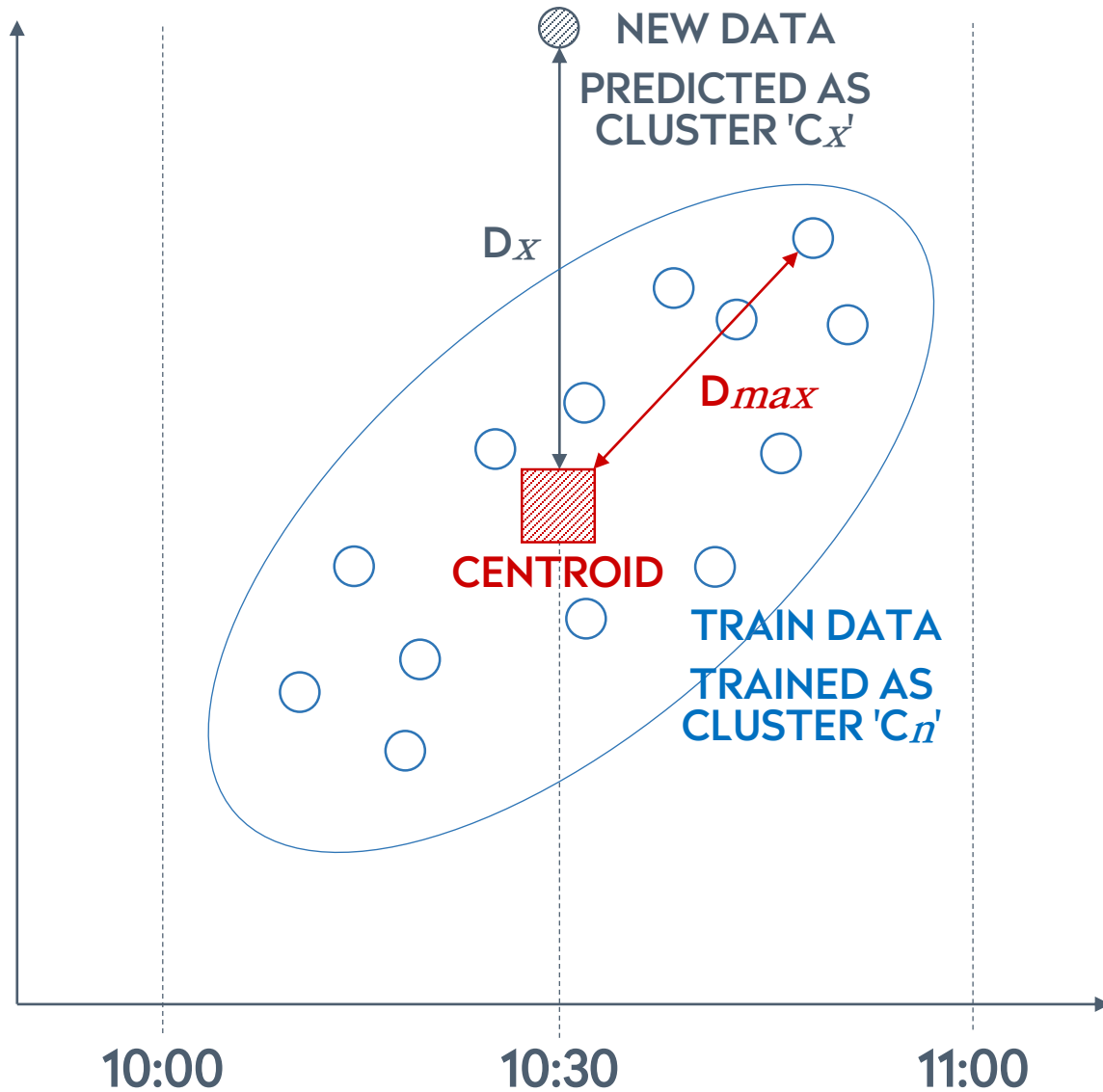
```
filter {  
  json {  
    source => "message"  
  }  
  mutate {  
    add_field => { "cpk" => "%{[_source][@fields][params][cpk]}" }  
    add_field => { "time" => "%{[_source][logdate]}" }  
    add_field => { "hit" => "%{[_source][@fields][params][hit]}" }  
    add_field => { "base_mck" => "%{[_source][@fields][params][base_mck]}" }  
    convert => {  
      "hit" => "integer"  
    }  
  }  
  prune {  
    whitelist_names => ["cpk", "hit", "base_mck", "time"]  
  }  
}
```



# PIPELINE



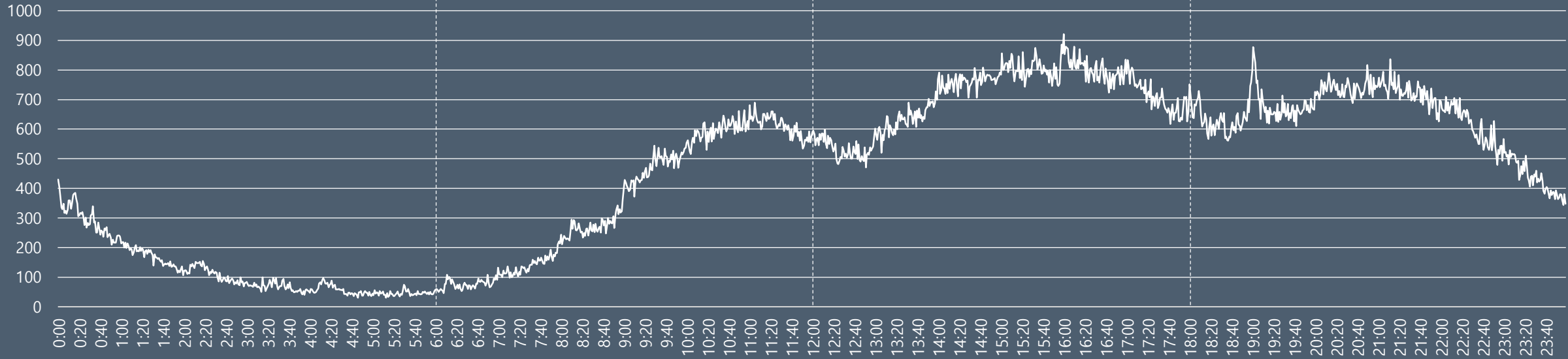
# KMEANS++





# EVALUATION

\* Validation metric: silhouette\_score



①  
하향 안정  
2 = 0.6122  
**3 = 0.6169**  
4 = 0.5617  
...  
→ k = 3

②  
상향  
2 = 0.6356  
**3 = 0.6432**  
4 = 0.5817  
...  
→ k = 3

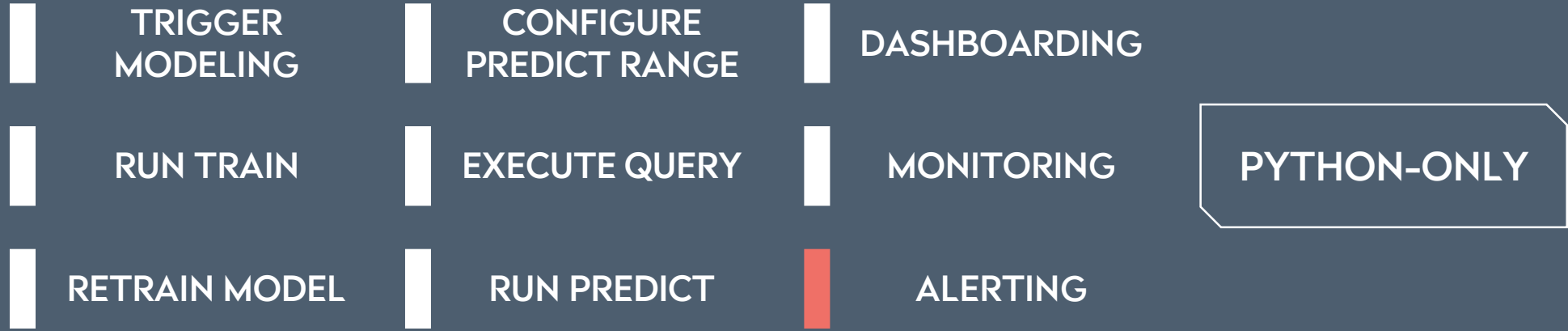
③  
상향 안정  
2 = 0.6191  
**3 = 0.6215**  
4 = 0.5692  
...  
→ k = 3

④  
하향  
2 = 0.6275  
**3 = 0.6347**  
4 = 0.5854  
...  
→ k = 3

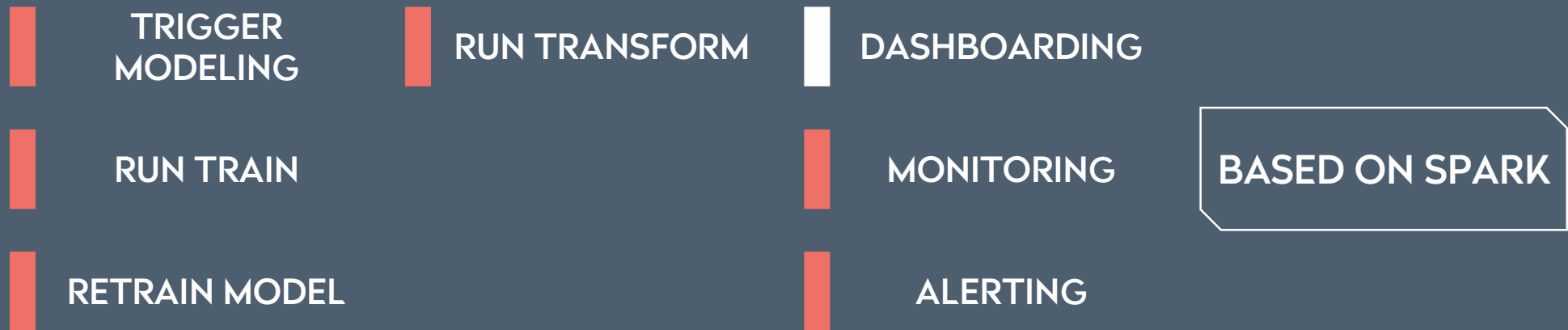
*Before we  
implemented Spark*

```
from keras.layers.recurrent import SimpleRNN
from keras.layers.recurrent import LSTM
from keras.layers.recurrent import GRU
from alibi_detect.od import OutlierProphet
from alibi_detect.od import OutlierSeq2Seq
from pyod.models import OCSVM
from pyod.models import LOF
from pyod.models import CBLOF
from pyod.models import HBOS
from pyod.models import KNN
from pyod.models import AvgKNN
from pyod.models import IForest
from pyod.models import XGBOD
from rrcf import RCTree
```

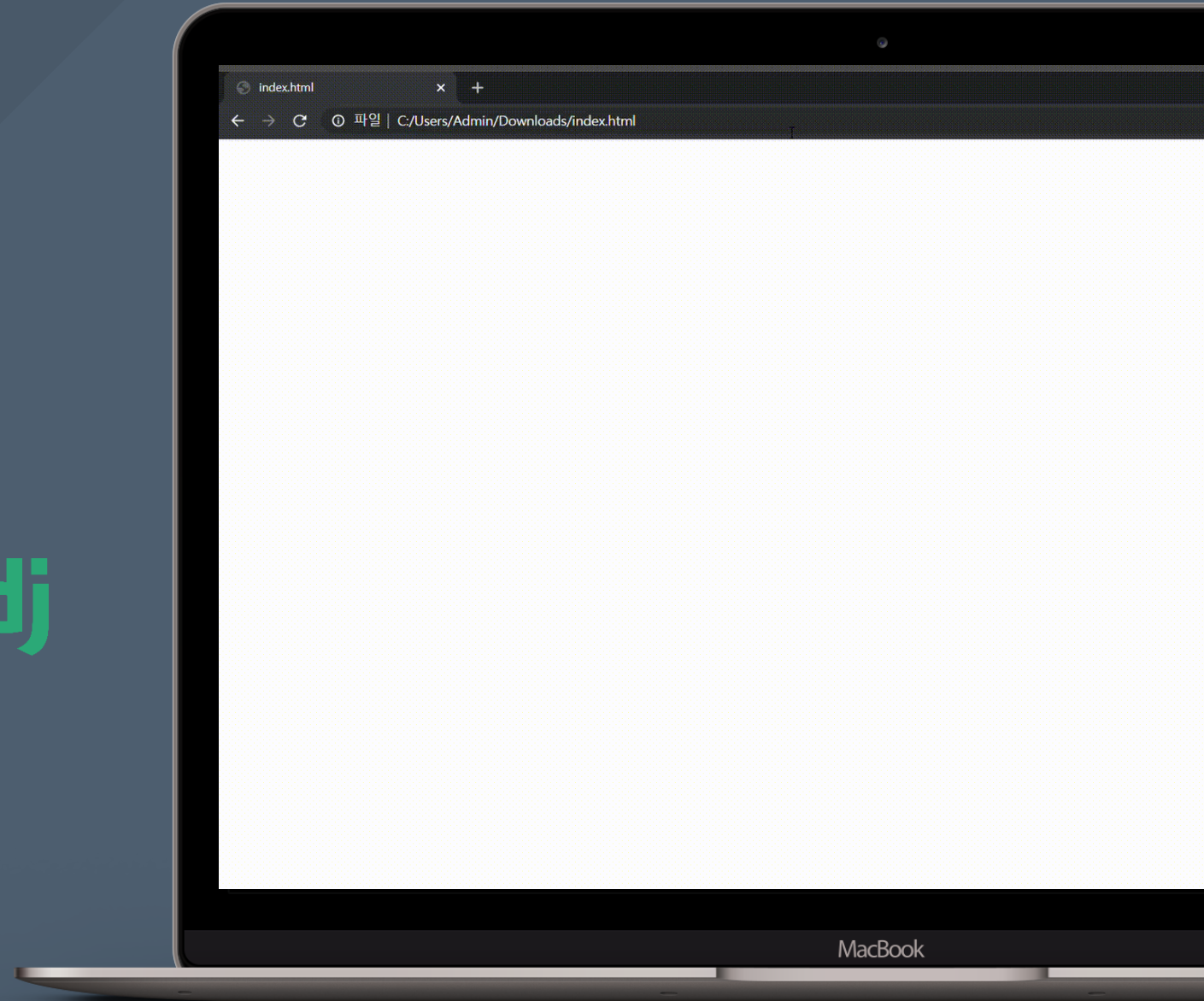
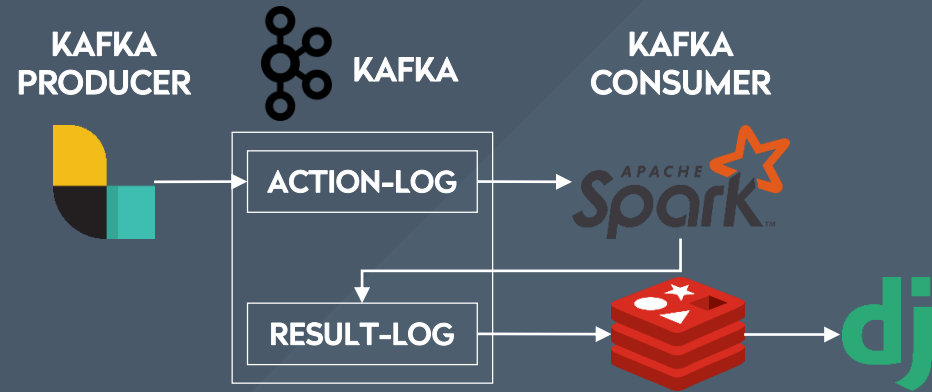




■ AUTOMATED TASK  
■ MANUAL TASK



# PIPELINE





## # KAFKA to SPARK

```
df = spark.readStream.format("kafka") \
    .option("kafka.bootstrap.servers", "<ip-address>:9092") \
    .option("subscribe", "action-log") \
    .option("startingOffsets", "latest") \
    .load()
```

## # AGGREGATION

```
streamingDF = action_detail_df \
    .withWatermark("timestamp", "10 seconds") \
    .groupBy(
        window("timestamp", "10 seconds"),
        action_detail_df.cpk, action_detail_df.base_mck) \
    .agg(count("*")) \
    .select("cpk", "base_mck", col("count(1)"), "window")
```

## # SPARK to KAFKA

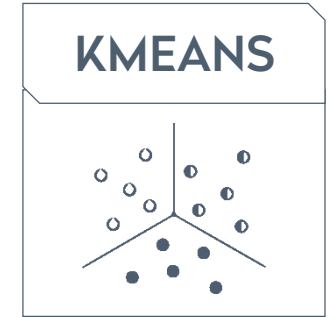
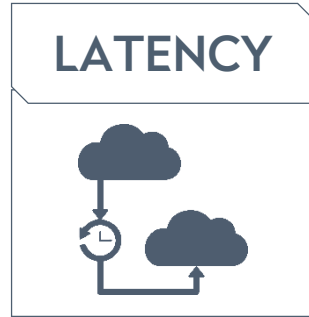
```
kafka_stream_dataframe = streamingDF \
    .selectExpr("CAST(cpk AS STRING) AS key", "to_json(struct(*)) AS value") \
    .writeStream.trigger(processingTime='10 seconds') \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "<ip-address>:9092") \
    .option("topic", "result-log") \
    .outputMode("update") \
    .option("checkpointLocation", "./tmp/checkpoint") \
    .start()
kafka_stream_dataframe.awaitTermination()
```



# REVIEW

소감  
시연



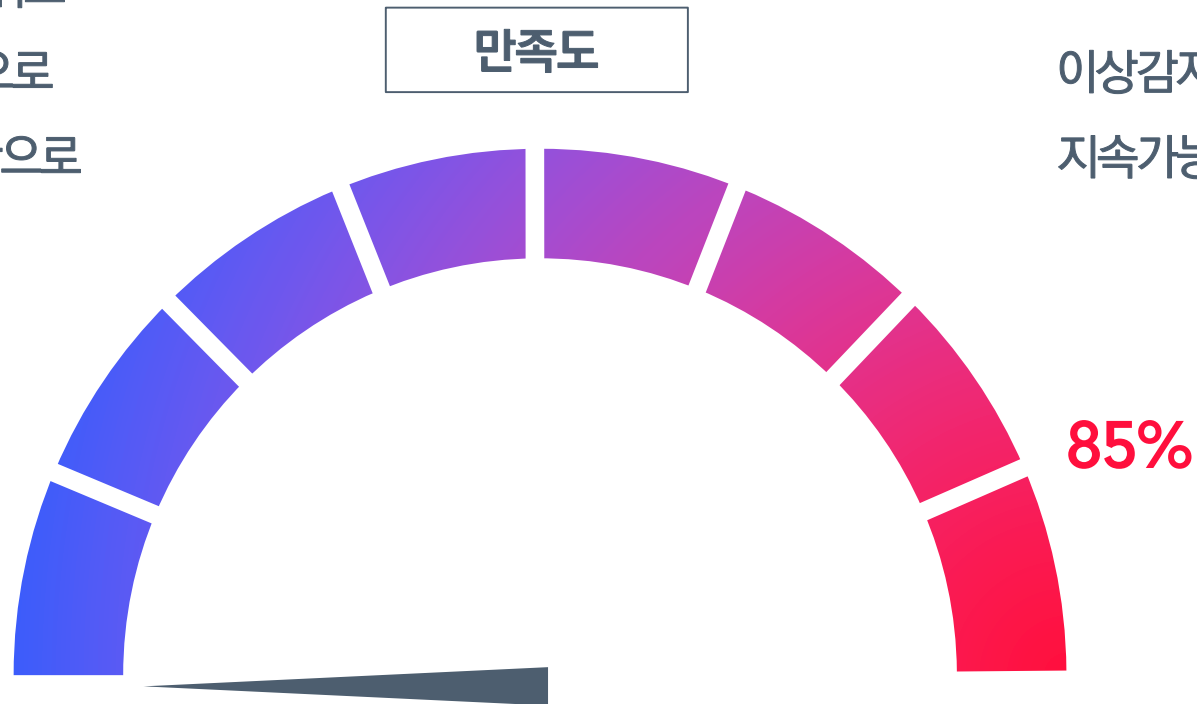


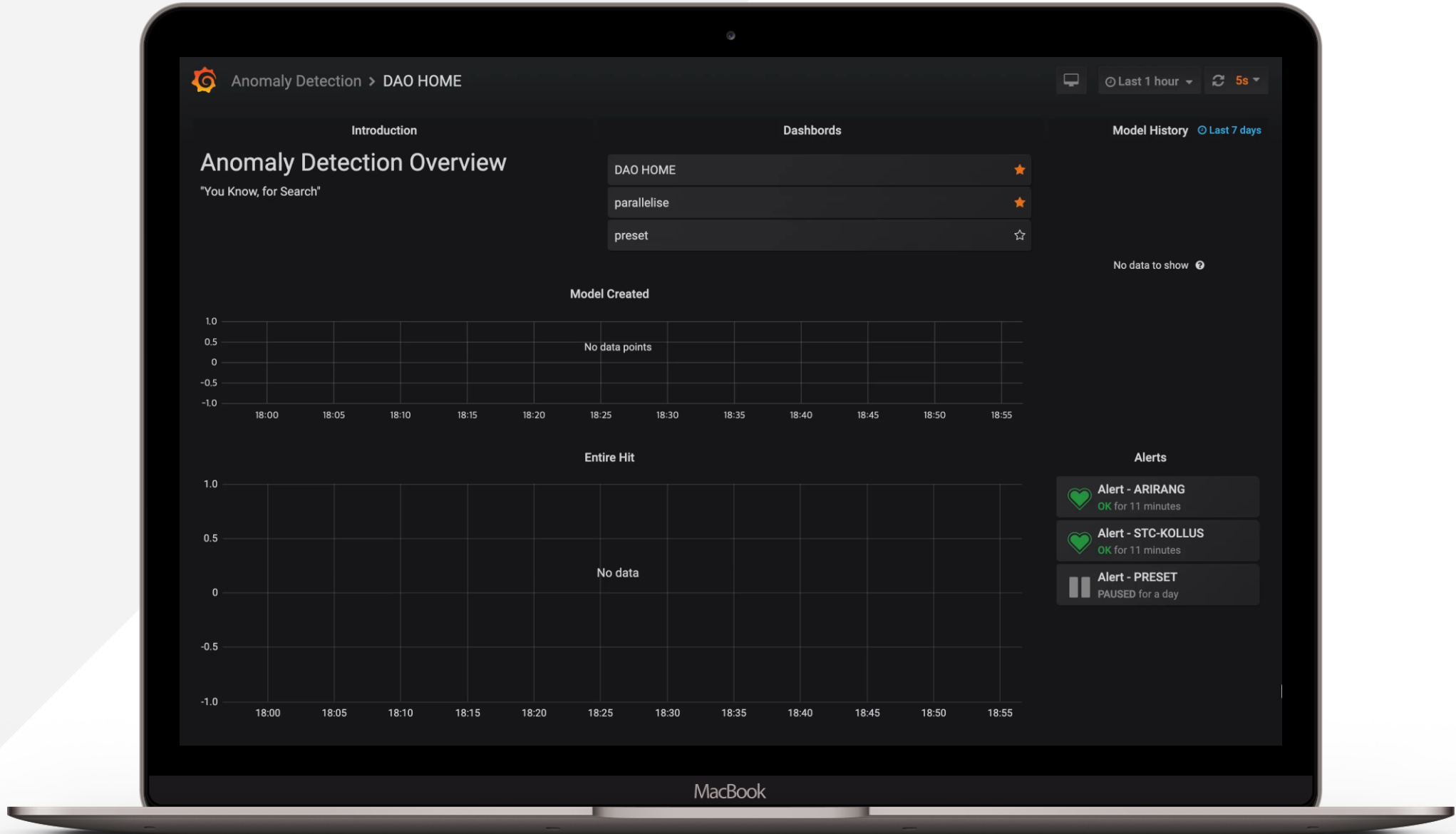
## GET THROUGH

- ▶ 한정된 자원, 닫힌 네트워크
- ▶ 처리 지연 없이 실시간으로
- ▶ 어떤 상황에서도 실시간으로

## FURTHER BEYOND

- 수치형 지표 고안 ◀
- 이상감지 활용 범위 확대 ◀
- 지속가능성, 안정성 확보 ◀







# THANK YOU FOR WATCHING



Except where otherwise noted, content on this slide is licensed under a Creative Commons Attribution 4.0 International license.  
Generated by Joohyun Keem.



본 문서는 16:9 FHD(1920x1080) 환경에 최적화되어 있습니다.