

# 비즈니스 혁명을 위한 빅데이터 플랫폼 및 활용사례



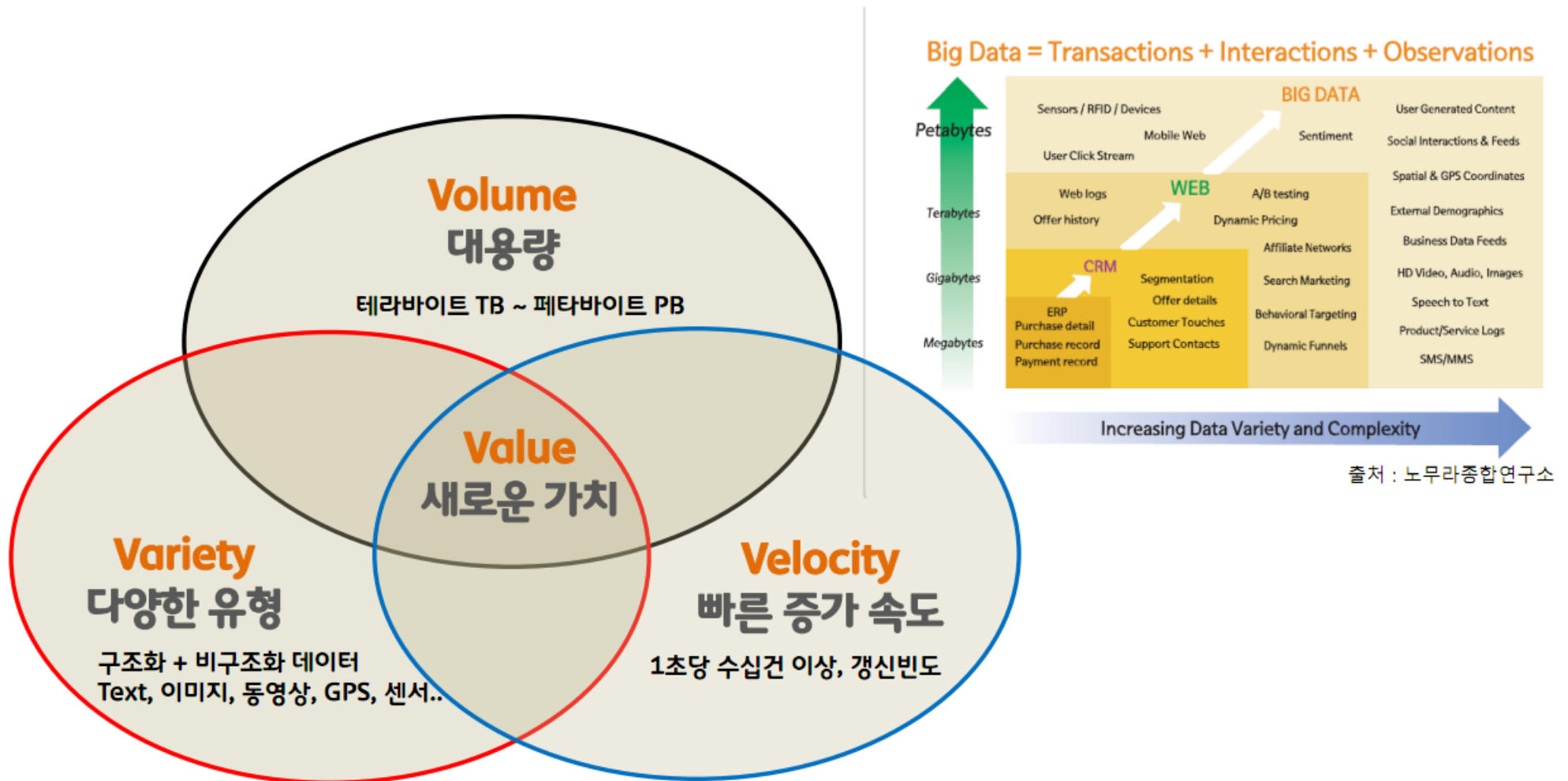
# 목 차

비즈니스 혁명을 위한 빅데이터 플랫폼 및 활용사례

- 1 ▶▶ 빅데이터와 비정형데이터
- 2 ▶▶ 빅데이터 플랫폼
- 3 ▶▶ 빅데이터 활용 사례

# 빅데이터 정의

조직 내부 및 외부에 존재하며 대용량(**Volume**)이고 다양한 유형(**Variety**) 데이터이며, 증가 속도가 빠르고(**Velocity**), 숨겨진 새로운 가치(**Value**)를 제공하는 특성을 갖습니다.



# 비정형데이터 증가

- 트위터, 페이스북, 유튜브, 인스타그램 등 소셜미디어(SNS), 포털에 기인한 비정형 데이터가 증가
  - 2015년 스마트폰 보급률 83%(세계 4위), 1인당 데이터 이용량 3GB/월 돌파
  - 그림, 영상, 문서, 의료기록, 음성정보 처럼 형태가 구조화되지 않은 데이터를 말함
  - 기업내 각종 문서, 홈페이지, 고객의 소리, 고객센터 상담메모, 사이버 상담 자료 등
  - 이메일, 트위터, 블로그처럼 모바일 기기와 온라인에서 생성되는 데이터
  - 일정한 규격이나 형태를 지닌 숫자데이터가 아님



자료 : 마이크로스트래티지



# 비정형 데이터 비율

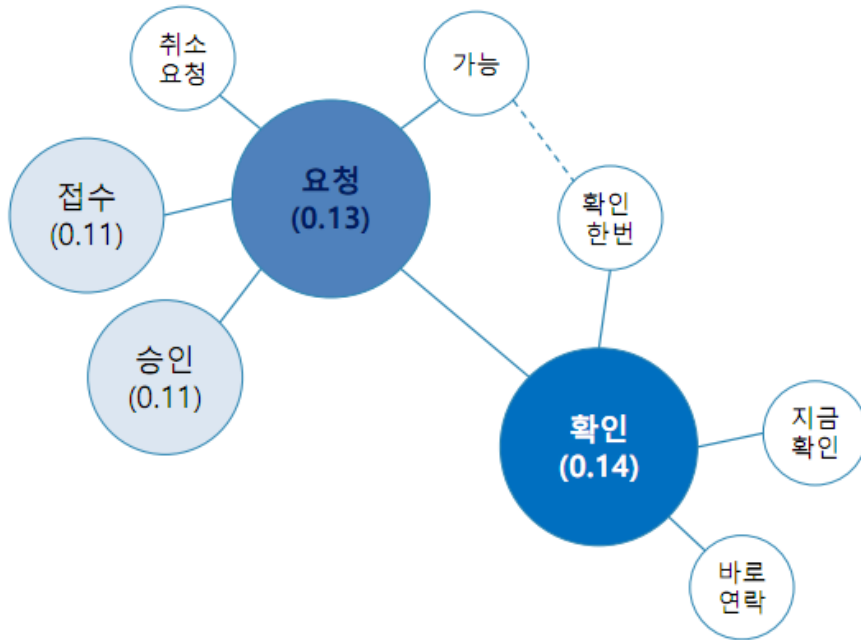
- 내부에 숨어 있는 비정형 데이터를 발굴



# 정형 + 비정형 데이터를 결합

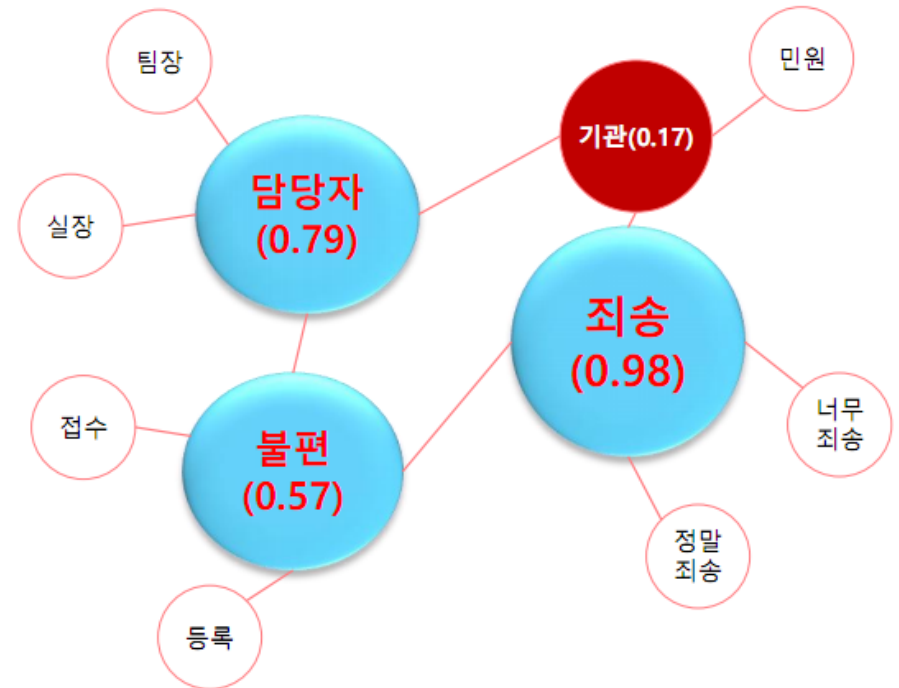
- 정형 + 비정형 데이터를 결합하여 의사결정을 지원

고객의 전체 키워드



타겟 고객 전체를 대상으로 키워드 분석하는 경우  
특징적인 키워드 추출이 어려움

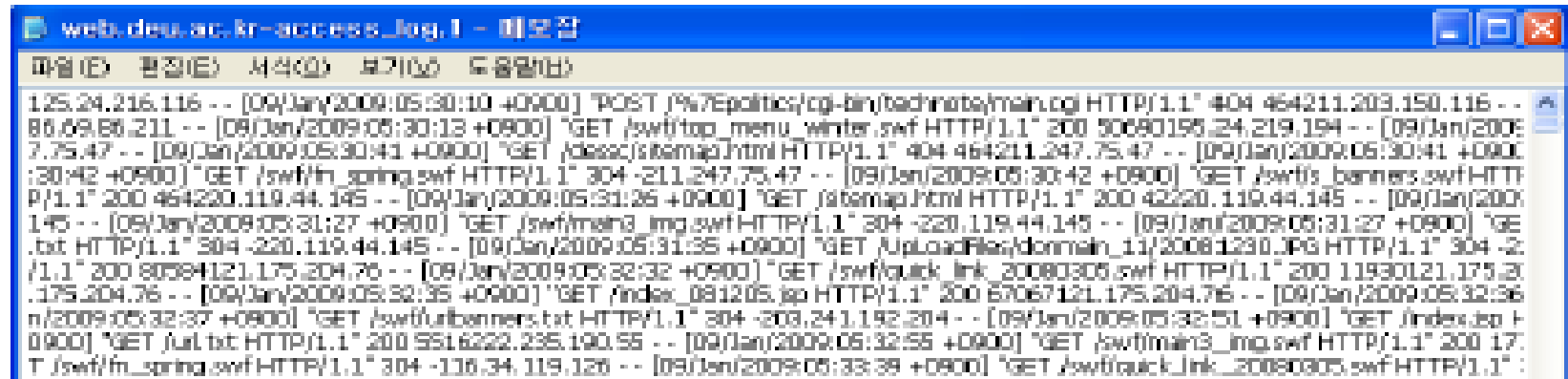
'기관을 언급한 고객'의 특징적 키워드



고객 특성에 따라 키워드 분석시  
“죄송”, “불편”, “담당자”, “금감원” 등의  
직접적이고 감정적인 키워드가 추출됨

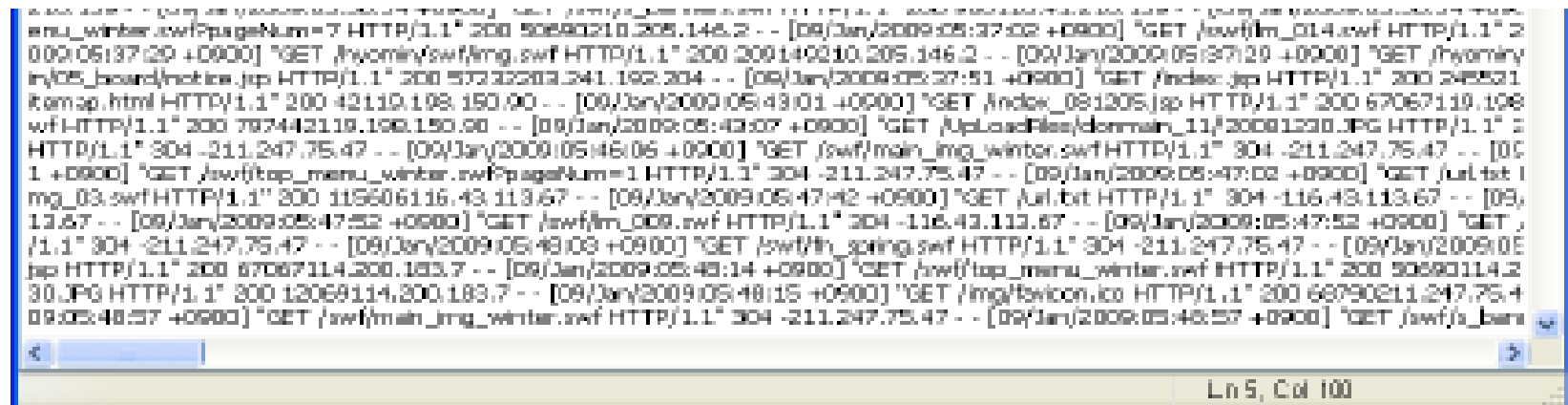
# 비정형 데이터 - 웹로그 데이터

- 웹로그 데이터



```
125.24.216.116 -- [09/Jun/2009:05:30:10 +0900] "POST /%7Epolicies/cgi-bin/technotes/main.cgi HTTP/1.1" 404 464211.203.150.116 --  
86.69.66.211 -- [09/Jun/2009:05:30:13 +0900] "GET /swf/top_menu_winter.swf HTTP/1.1" 200 50680150.24.219.194 -- [09/Jun/2009:  
7.75.47 -- [09/Jun/2009:05:30:41 +0900] "GET /dcssc/sitemap.html HTTP/1.1" 404 464211.247.75.47 -- [09/Jun/2009:05:30:41 +0900:  
:30:42 +0900] "GET /swf/tn_spring.swf HTTP/1.1" 304 -211.247.75.47 -- [09/Jun/2009:05:30:42 +0900] "GET /swf/s_banners.swf HTTP  
P/1.1" 200 464220.119.44.145 -- [09/Jun/2009:05:31:26 +0900] "GET /sitemap.html HTTP/1.1" 200 42220.119.44.145 -- [09/Jun/2009:  
145 -- [09/Jun/2009:05:31:27 +0900] "GET /swf/main3_img.swf HTTP/1.1" 304 -220.119.44.145 -- [09/Jun/2009:05:31:27 +0900] "GE  
.txt HTTP/1.1" 304 -220.119.44.145 -- [09/Jun/2009:05:31:35 +0900] "GET /UploadFiles/domain_11/20081230.JPG HTTP/1.1" 304 -2:  
/1.1" 200 80384121.175.204.76 -- [09/Jun/2009:05:32:32 +0900] "GET /swf/quick_link_20080305.swf HTTP/1.1" 200 11530121.175.2:  
.175.204.76 -- [09/Jun/2009:05:32:35 +0900] "GET /index_081205.jsp HTTP/1.1" 200 67067121.175.204.76 -- [09/Jun/2009:05:32:36  
n/2009:05:32:37 +0900] "GET /swf/turbanners.txt HTTP/1.1" 304 -203.241.192.204 -- [09/Jun/2009:05:32:51 +0900] "GET /index.jsp +  
0900] "GET /url.txt HTTP/1.1" 200 5516222.235.190.55 -- [09/Jun/2009:05:32:55 +0900] "GET /swf/main3_img.swf HTTP/1.1" 200 17:  
T /swf/tn_spring.swf HTTP/1.1" 304 -116.34.119.120 -- [09/Jun/2009:05:33:38 +0900] "GET /swf/quick_link_20080305.swf HTTP/1.1" :
```

211.76.351.28 - admin [13/Jun/2015:13:10:35 +0900] "GET /  
HTTP/1.1" 200



```
enu_winter.swf?pageNum=7 HTTP/1.1" 200 50680210.205.146.2 -- [09/Jun/2009:05:37:02 +0900] "GET /swf/tn_014.swf HTTP/1.1" 2  
009:05:37:29 +0900] "GET /hyominy/swf/img.swf HTTP/1.1" 200 209149210.205.146.2 -- [09/Jun/2009:05:37:29 +0900] "GET /hyominy  
in/05_board/notice.jsp HTTP/1.1" 200 57322202.241.192.204 -- [09/Jun/2009:05:37:51 +0900] "GET /index.jsp HTTP/1.1" 200 245531  
kemap.html HTTP/1.1" 200 42110.108.190.90 -- [09/Jun/2009:05:43:01 +0900] "GET /index_081205.jsp HTTP/1.1" 200 67067110.108  
wf HTTP/1.1" 200 797442110.190.150.90 -- [09/Jun/2009:05:43:07 +0900] "GET /UploadFiles/domain_11/20081230.JPG HTTP/1.1" 2  
HTTP/1.1" 304 -211.247.75.47 -- [09/Jun/2009:05:46:06 +0900] "GET /swf/main_img_winter.swf HTTP/1.1" 304 -211.247.75.47 -- [09:  
1 +0900] "GET /swf/top_menu_winter.swf?pageNum=1 HTTP/1.1" 304 -211.247.75.47 -- [09/Jun/2009:05:47:02 +0900] "GET /url.txt  
mg_03.swf HTTP/1.1" 200 115606116.43.113.67 -- [09/Jun/2009:05:47:42 +0900] "GET /url.txt HTTP/1.1" 304 -116.43.113.67 -- [09:  
13.67 -- [09/Jun/2009:05:47:52 +0900] "GET /swf/tn_009.swf HTTP/1.1" 304 -116.43.113.67 -- [09/Jun/2009:05:47:52 +0900] "GET /  
/1.1" 304 -211.247.75.47 -- [09/Jun/2009:05:48:03 +0900] "GET /swf/tn_spring.swf HTTP/1.1" 304 -211.247.75.47 -- [09/Jun/2009:05:  
jsp HTTP/1.1" 200 67067114.200.183.7 -- [09/Jun/2009:05:48:14 +0900] "GET /swf/top_menu_winter.swf HTTP/1.1" 200 50680114.2  
30.JPG HTTP/1.1" 200 12068114.200.183.7 -- [09/Jun/2009:05:48:15 +0900] "GET /img/bvicon.co HTTP/1.1" 200 66790211.247.75.4  
09:05:48:57 +0900] "GET /swf/main_img_winter.swf HTTP/1.1" 304 -211.247.75.47 -- [09/Jun/2009:05:48:57 +0900] "GET /swf/s_bars
```

# 비정형 데이터 - 트위터 데이터

- 트위터 검색 결과 및 사용자 정보 결과 데이터

```
{
  "completed_in":0.031,
  "max_id":122078461840982016,
  "max_id_str":"122078461840982016",
  "next_page":"?page=2&max_id=122078461840982016&q=blue%20angels&rpp=5",
  "page":1,
  "query":"blue+angels",
  "refresh_url":"?since_id=122078461840982016&q=blue%20angels",
  "results":[
    {
      "created_at":"Thu, 06 Oct 2011 19:36:17 +0000",
      "entities":{
        "urls":{
          {
            "url":"http://t.co/L9JXJ2ee",
            "expanded_url":"http://bit.ly/q9fyz9",
            "display_url":"bit.ly/q9fyz9",
            "indices":{
              37,
              57
            }
          }
        ]
      },
      "from_user":"SFist",
      "from_user_id":14093707,
      "from_user_id_str":"14093707",
      "geo":null,
      "id":122032448266698752,
      "id_str":"122032448266698752",
      "iso_language_code":"en",
      "metadata":{
        "recent_retweets":3,
        "result_type":"popular"
      },
      "profile_image_url":"http://a3.twimg.com/profile_images/51584619/SFist07_normal.jpg",
      "source":"<a href='\"http://twitter.com/tweetbutton\"'>Tweet Button</a>",
      "text":"Reminder: Blue Angels practice today http://t.co/L9JXJ2ee",
      "to_user_id":null,
    }
  ]
}
```

```
[
  {
    "name": "Twitter API",
    "profile_sidebar_border_color": "87bc44",
    "profile_background_tile": false,
    "profile_sidebar_fill_color": "e0ff92",
    "location": "San Francisco, CA",
    "profile_image_url": "http://a3.twimg.com/profile_images/689684365/api_normal.png",
    "created_at": "Wed May 23 06:01:13 +0000 2007",
    "profile_link_color": "0000ff",
    "favourites_count": 2,
    "url": "http://apiwiki.twitter.com",
    "contributors_enabled": true,
    "utc_offset": -28800,
    "id": 6253282,
    "profile_use_background_image": true,
    "profile_text_color": "000000",
    "protected": false,
    "followers_count": 160752,
    "lang": "en",
    "verified": true,
    "profile_background_color": "c1dfee",
    "geo_enabled": true,
    "notifications": false,
    "description": "The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
    "time_zone": "Pacific Time (US & Canada)",
    "friends_count": 19,
    "statuses_count": 1858,
    "profile_background_image_url": "http://a3.twimg.com/profile_background_images/59931895/twitterapi-background-new.png",
    "status": {
      "coordinates": null,
      "favorited": false,
      "created_at": "Tue Jun 22 16:53:28 +0000 2010",
      "truncated": false,
      "text": "@Demonicpagan possible some part of your signature generation is incorrect & fails for real reasons.. follow up on the list if you suspect",
      "contributors": null,
      "id": 16783999399,
      "geo": null,
      "in_reply_to_user_id": 6339722,
    }
  }
]
```

# 비정형 데이터 - 생체신호

- 빅데이터의 모습 - 생체신호(심박수와 열화상 센서 데이터)

body1(심박수) (2) - 메모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)									
0	1	2	3	4	5	6	7	8	9	10	11	12	13
.9	74.6	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
6	74.6	74.6	74.4	74.6	74.6	74.4	74.4	74.4	74.4	74.4	74.4	74.4	74.4
4.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6
.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
73.6	73.6	8	74.4	74.6	74.4	74.6	74.4	74.6	74.4	74.6	74.4	74.6	74.4
6	74.6	74.6	74.6	74.6	74.6	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
1	80.8	80.6	79.9	79.1	78.4	77.2	76.6	76.6	76.6	76.6	76.6	76.6	76.6
9	74.6	74.6	74.6	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
74.6	74.6	74.6	74.6	74.6	74.6	74.9	74.6	74.6	74.6	74.6	74.6	74.6	74.6
74.9	74.9	74.9	74.9	75.2	74.9	75.2	75.2	75.2	75.2	75.2	75.2	75.2	75.2
76.1	75.7	75.7	75.7	75.4	75.2	75.2	75.2	75.2	75.2	75.2	75.2	75.2	75.2
74.6	74.6	74.6	74.4	74.4	74.4	74.1	74.1	74.1	74.1	74.1	74.1	74.1	74.1
4.6	74.6	74.9	74.9	74.6	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
4.7	85.3	86.1	86.8	87.4	86.6	86.1	85.5	85.5	85.5	85.5	85.5	85.5	85.5
4.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
22	74.6	74.6	74.6	74.6	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
.2	89.7	89.2	91.2	91.2	89.7	88.7	88.7	88.7	88.7	88.7	88.7	88.7	88.7
.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9	74.9
6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6	74.6
0	82.9	83.3	84.2	85.1	85.6	86.1	86.1	86.1	86.1	86.1	86.1	86.1	86.1

ThermalDat(열화상) - 메모장

파일(F)	편집(E)	서식(O)	보기(V)	도움말(H)														
76800	24.49	24.49	24.01	23.39	23.52	23.82	23.60	23.34	23.52	23.47	23.39	23.69	23.79	24.01	23.56	23.87	24.08	24.08
4	32.78	32.98	32.71	32.68	32.47	32.66	32.61	32.22	32.26	32.22	32.13	31.90	31.13	28.87	27.91	27.15	26.26	26.34
24.61	24.32	23.39	24.01	23.82	23.60	23.26	23.69	23.56	23.56	23.56	23.96	23.79	23.87	23.87	24.21	24.21	24.36	24.36
7	32.71	33.17	32.87	32.57	32.84	32.54	32.38	31.90	32.22	32.13	32.10	31.77	30.61	28.71	28.25	27.43	26.93	26.59
23.69	24.01	23.73	23.96	23.26	23.87	23.12	23.69	23.47	23.96	23.39	24.13	23.30	23.91	23.65	24.28	23.96	24.17	23.96
8	32.87	32.71	32.84	32.47	32.61	32.26	32.26	31.98	31.98	32.10	32.10	32.18	31.45	29.40	28.63	27.87	27.79	27.02
24.40	23.47	23.79	23.79	23.65	23.30	23.79	23.82	23.79	24.13	23.56	23.91	24.17	24.02	24.17	24.21	24.21	24.13	23.96
8	32.71	33.00	32.78	32.61	32.87	32.68	32.30	32.10	32.26	32.22	32.01	31.98	32.22	30.61	29.23	28.76	28.16	27.75
24.32	23.82	23.39	23.52	23.82	23.73	23.34	23.82	24.01	23.43	24.02	23.73	23.96	24.36	24.57	23.87	24.66	24.13	24.13
4	32.90	32.78	32.89	32.81	32.71	32.66	32.34	32.10	32.14	32.13	32.26	32.22	31.94	31.77	30.21	29.44	28.59	27.72
23.91	24.32	24.01	23.96	23.65	23.96	23.43	23.87	23.56	23.96	23.91	24.01	24.17	24.40	24.02	24.28	24.49	24.32	24.32
4	33.36	32.87	32.89	32.98	32.72	32.68	32.54	32.18	31.98	32.13	32.18	32.18	32.05	31.94	30.97	29.68	28.49	27.43
24.61	24.66	23.79	23.79	23.47	23.82	23.65	23.69	23.79	24.40	24.32	24.36	24.28	24.57	23.96	24.08	24.44	24.61	24.61
6	34.08	32.87	32.90	32.71	32.81	32.57	32.57	32.18	32.13	32.18	32.34	32.22	32.05	31.81	31.81	28.98	28.00	27.11
24.40	24.32	23.96	24.17	23.91	24.02	23.79	24.17	24.02	23.65	24.44	24.01	24.32	24.17	24.08	24.36	24.32	24.53	24.53
0	34.60	33.44	33.06	32.68	32.89	32.47	32.38	32.22	32.30	31.98	32.13	32.18	32.18	32.01	31.94	29.86	27.79	27.15
24.08	24.40	23.96	24.28	23.52	24.01	23.82	23.96	23.79	24.32	24.08	24.71	24.13	24.28	24.57	24.36	24.01	24.28	24.28
6	34.70	34.11	32.98	33.03	32.90	32.66	32.61	32.38	32.54	31.85	32.30	32.34	32.26	32.05	32.14	31.13	28.04	26.98
23.96	24.08	24.08	24.13	23.43	24.02	23.21	24.02	24.13	23.91	23.87	24.36	23.82	24.21	23.96	24.21	24.28	24.53	24.53
5	34.61	34.51	33.42	32.87	33.00	32.57	32.57	32.22	32.47	32.14	32.18	32.13	32.01	31.85	32.18	31.09	29.36	27.28
24.40	24.01	24.01	23.82	23.79	24.02	23.91	24.08	23.69	24.40	23.87	24.17	24.28	24.44	24.32	24.01	24.36	24.57	24.57
6	34.41	34.60	34.30	32.71	32.61	32.57	32.54	32.05	32.34	32.42	32.51	32.18	32.10	32.14	32.18	32.01	30.97	29.36
24.32	23.82	23.96	23.69	23.69	24.01	23.96	23.91	24.08	24.08	24.21	23.87	24.28	24.53	23.73	24.40	24.44	24.61	24.61
3	34.23	34.50	34.55	33.42	32.51	32.51	32.47	32.22	32.30	32.26	32.43	32.34	32.14	32.05	32.14	32.05	31.94	32.68



# 비정형 데이터 인식수준

- 정답(속기) 내용과 텍스트 변환 결과를 비교한 후 일치하는 단어는 정인식(正認識), 다르게 인식한 결과는 오인식(誤認識), 텍스트 결과가 없는 단어는 미인식(未認識)으로 산정한 결과, 평균 80% 수준 변환

## 1 속기사 (전사)

따뜻한 금융 을 실천 하는 신한카드 상담원      입니다 고객님의 무엇 을 도와 드릴까요  
아 네 안녕 하셨습니다 제가 신용 카드 를 신청 했는데요  
네 예 그거 받는 곳 을 직장 으로 했는데 직장 에 그 시간 이 지나도 좀 안 와가지고 못받  
았거든요 (소리) 다른 사람 이 그거 받았나 싶어 가지구요 제 집 으로 다시 보내주시면 안  
될까요 네 집 로 받아보시고 싶으시다구요 네 네 네  
네 예 제가 확인 도와 드릴 텐데 잠시 만 기다려 주시겠습니까  
네 네 고객님의 제가 확인 중 에 있습니다 실례 지만 고객님의 성함 이 어떻게 되십니까  
네 고객님의 자택 주소 가 어떻게 되십니까  
네 부산 시 동래 구 낙민 동 중앙 하이츠  
네 소중 한 정보 확인 감사 합니다 네 고객님의 지금 현재 아 확인 해 본 결과 해당 카드 는  
지금 배송 업체 에 있으세요  
아 (소리) 배송 업체 아 그럼 아직 직장 안 온거네요 그러면 직장 그대로 해 놓을게요  
아 고객님의 아직 연락 을 못 받으 셸어요  
아 카드 온다는 연락이요 네  
네 네 네 저 혹시 연락 가능 하신 연락처 가 어떻게 등록 이 되어있는지 (소리) 아 확인 도  
와 드릴 텐데요 네 잠시 만 기다려 주시겠습니까  
네 네 고객님의 어 현재 그러면 자택 이랑 직장 이랑 거리 가 많이 (소리)  
네 네 한 시간 정도 걸려요 한 시간 정도 걸리세요  
네 (잡음)  
네 고객님의 제가 확인 도와 드릴 텐데 잠시 만 기다려 주시겠습니까  
네 네 고객님의 그러면 저희 가 발송 (소리) 아 발송 업체 에 연락 을 해서 아 (소리) 자택 주  
소 로 받아보실 수 있게 (소리) 어 도와 드릴 텐데요  
네 주소 받으실 네 주소 받으실 자택 주소 가 어떻게 되십니까  
아 (소리) 등록 된 건 똑 같아요  
네 아 다시 한 번 확인 더 부탁드립니다  
예 부산 시 동래 구 낙민 동 중앙 하이츠  
네 지금 직장 주소 로 되어 (소리) 있는 곳 에서 자택으로 (소리) 받아보시기 원하신다고요  
네 네  
네 저 고객님의 요청 하신대로 제가 아 (소리) 메모 전달 해 드릴 텐데요 고객님의 다른 더 문  
의 사항 은 (소리) 없으십니까 (잡음)  
예 예 없습니다 네 (소리)  
네 감사 합니다 신한카드 상담원 (      )였습니다 즐거운 하루 되십시오  
네 감사 합니다  
네 감사 합니다 (잡음)

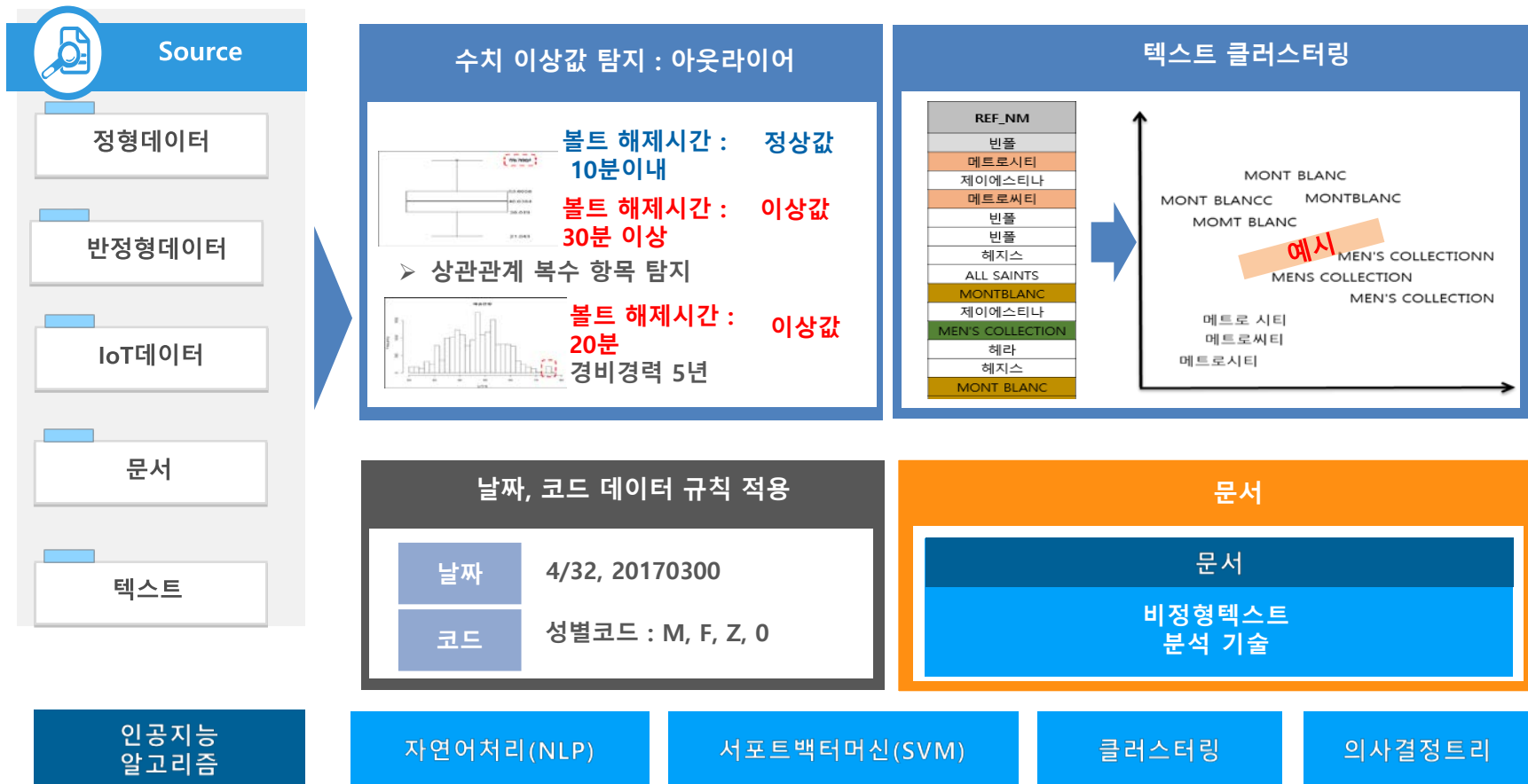
## 2 음성인식 (STT)

오인식      미인식

따뜻한/ 금융/ 을/ 실천/ 하는/ 신한카드/ 상담원/      / 입니다/ 고객님의/ 무엇 을/ 도  
와/ 드릴까요/ 아/ 네/ 안녕/ 그럼 하셨습니다/ 제가/ 신용/ 카드/ 를/ 실-문(신청) 되나(했  
는데요) 네/ (예) 그래서(그거) 당면(받는) 곳/ 온(을) 직장/ 으로 돼 왔는데(했는데) 직장  
에 그/ 시간/ 이/ 지나도/ 좀/ 가용(안) 하자-거(와가지고) 못 받았거든요/ (혹시) 네/ 다  
른/ 사람/ 이/ 그거/ 받았나/ 싶어/ 가지구요/ 제 집/ 으로/ 다시/ 보내주시면/ 안될까요  
네/ 집/ 으로/ 받아보시고/ 싶으시다구요/ 네 네 (네) 네/ 예/ 제가/ 확인/ 도와/ 드릴/  
텐데/ 잠시/ 만/ 기다려/ 주시겠습니까/ 네/ 고객님의/ 제가/ 확인/ 중/ 예/ 있습니다/ 실례  
/ 지만/ 고객님의/ 성함/ 이/ 어떻게/ 되십니까/ 네/ 고객님의/ 자택/ 주소/ 가/ 어떻게/ 되십  
니까/ 부산/ 시/ 동래/ 구/ 낙(낙민) 원(동) 중(중앙) 했었(중) 하이츠/ 네/ 소중/ 한/ 정보/  
확인/ 감사/ 합니다/ [근데] 고객님의/ 지금/ 현재/ 아/ 확인/ 해/ 본/ 결과/ 해당/ 카드/ 는/  
지금/ 배송/ 업체/ 에/ 있으세요/ 아/ 네/ 아 배송 업체 그럼/ 아직/ 직장/ 안/ 온/ 거  
네요/ 그러면/ 직장/ 그대로/ 해/ 놓은(놓을게요) 어/ 그럼(고객님) 다시(아직) 연락/ 을/  
못/ 받으셨어요/ 아/ 카드/ 온다는/ 연락/ 아요 네/ 네/ 네/ 네/ 저/ 혹시/ 연락/ 가능/  
하신/ 연락처/ 가/ 어떻게/ 등록/ 이/ 되어/ 있는지요 [아 확인 도와 드릴 텐데요] 네 잠  
시/ 만/ 기다려/ 주시겠습니까/ 네/ 고객님의/ 현재/ 어 그러면/ 자택/ 이랑/ 직장/ 이랑/  
거리/ 가/ 많이/ 뭐-왔으세요(머세요) 네/ 네 한/ 시간/ 정도/ 걸려요(걸리요) 한/ 시간/  
정도/ 걸리세요/ (네) 네/ 그럼(고객님) 제가/ 확인/ 도와/ 드릴/ 텐데/ 잠시/ 만/ 기다려/  
주시겠습니까/ 네/ 고객님의/ 그럼(그러면) 저희/ 가/ 발송/ 다음(반) 아/ 발송/ 업체/ 예/  
연락/ 을/ 해서/ 아/ [고객님 자택 주소로 받아 보실 수] 왔든(있게) 도와/ 드릴/ 텐데요  
/ 네/ 꼭(주소) 받으실/ 네/ 주소/ 받으실/ 자택/ 주소/ 가/ 어떻게/ 되십니까/ (아) 근데  
(그) 행복(등록) 되는(된) 걸(건) 부탁(꼭같아요) [네 아 다시 한 번 더 확인 부탁 드리겠습  
니다] 예/ 부산/ 시/ 동래/ 구/ 낙-단(낙민) 등/ 중앙(중앙) (하이츠) 네/ 지금/ 직장/ 주  
소/ 로/ 되어/ 왔으면(있는) 곳/ 에서/ 자택/ 으로/ 받아/ 보시게(보시기) 원하신다고요/  
네/ (네) 아(저) 쪽으로(고객님) 요청/ 하신대로/ 제가/ 아/ 전답 메모/ 전달/ 해/ 드릴/ 텐  
데요/ 고객님의/ 다른/ 더/ 문의/ 사항/ 은/ 있으신(없으십니까) 있으십니까 예/ 예 방자한  
합니다(없습니다) 네 네 감사/ 합니다/ 신한카드/ 상담원/      /였습니다/ 즐거운/  
하루/ 되십시오/ 네/ 감사/ 합니다/ 네/ 감사/ 합니다/ ..

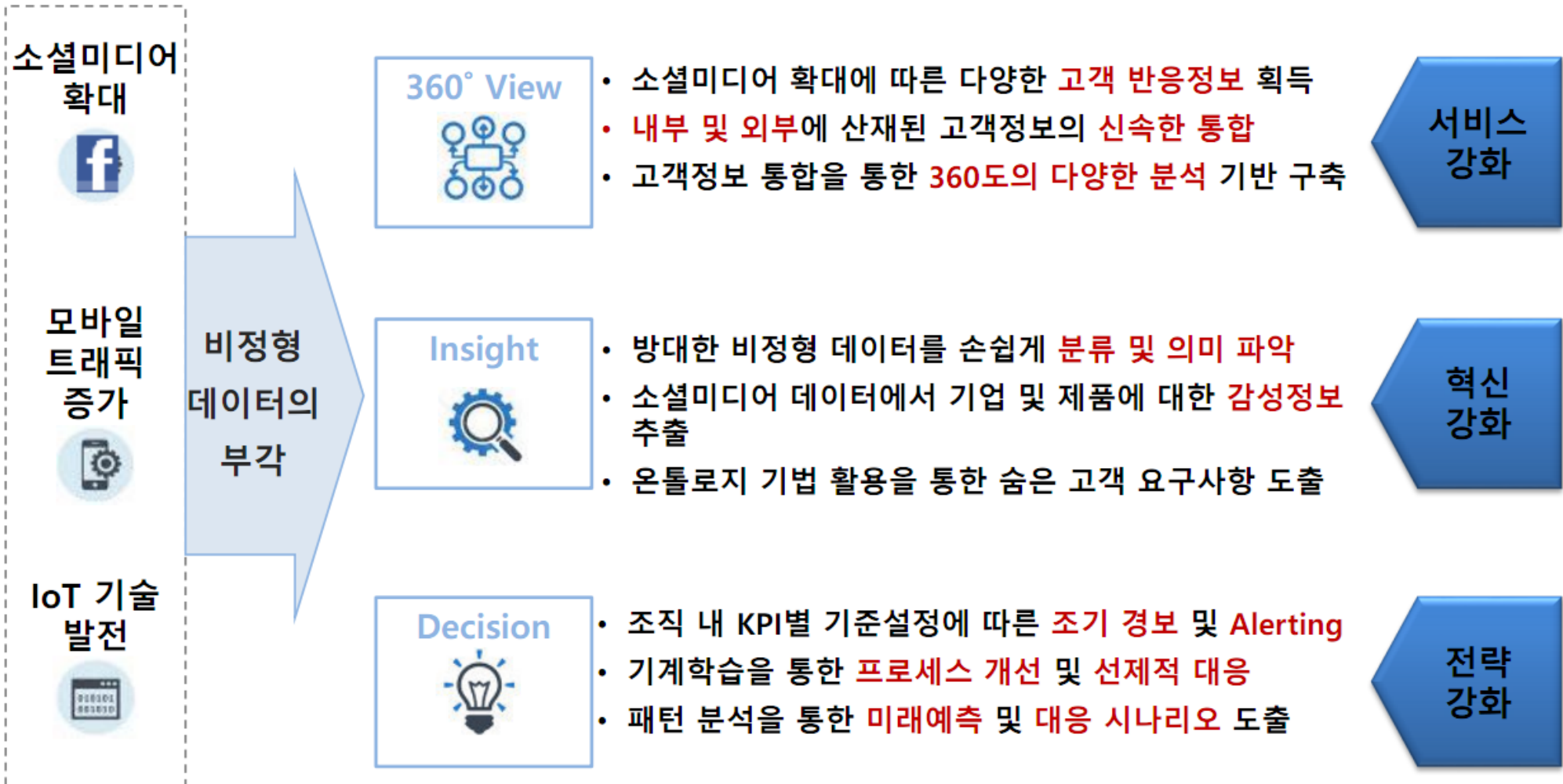
# 비정형 데이터 정제 · 정형화

- 데이터분포 분석을 통한 이상값 탐지
- 텍스트 데이터 클러스터링을 통한 데이터 정형화



# 빅데이터 분석 활용

- 빅데이터 분석을 통해 경영혁신에 활용함

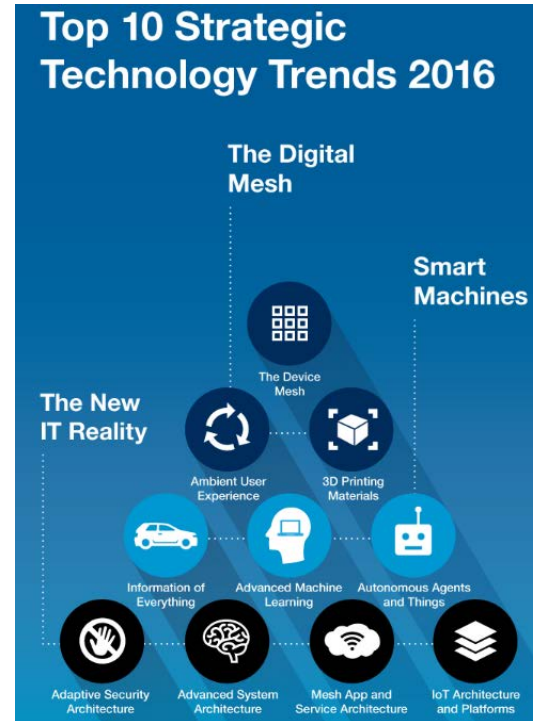


# 빅데이터 분석의 시대적 흐름

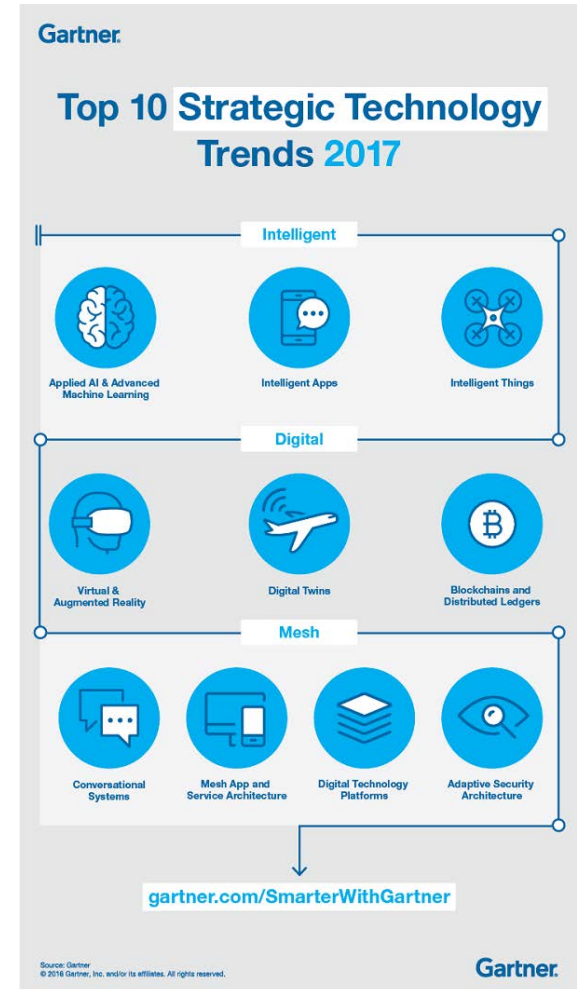
Gartner에서 발표하는 10대 전략 기술 동향에서 빅데이터의 위치, 중요성을 파악한다

빅데이터는 2010년대 초반 급부상 후 점차 사라짐

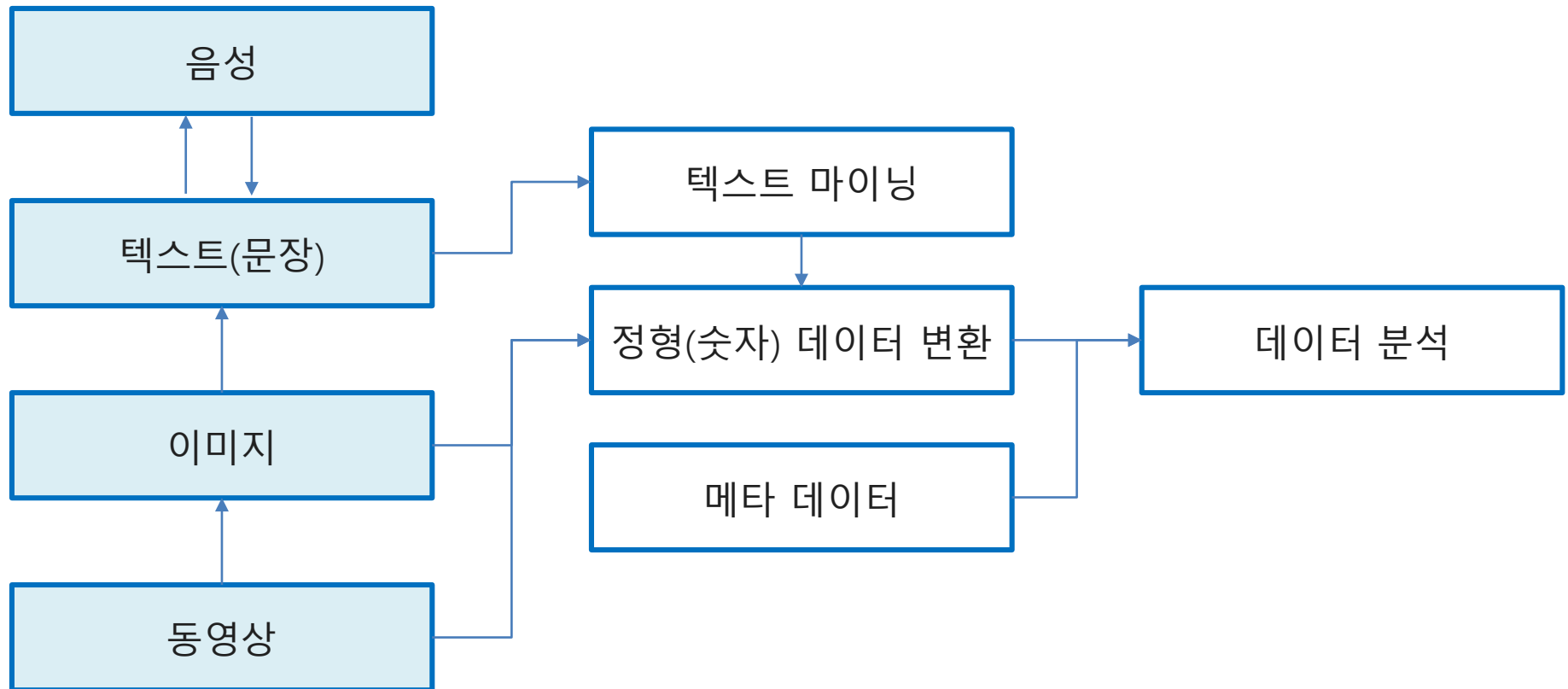
순위	2011년	2012년	2013년	2014년	2015년
1	클라우드 컴퓨팅	미디어 태블릿 그 이후	모바일 대전	다양한 모바일 기기관리	언제 어디서나 컴퓨팅 사용이 가능한 컴퓨팅 애플리케이션
2	모바일 앱과 미디어 태블릿	모바일 중심 애플리케이션과 인터페이스	모바일 앱 & HTML5	모바일 앱과 애플리케이션	사물인터넷
3	소셜 커뮤니케이션 및 협업	상황인식과 소셜이 결합된 사용자 경험	퍼스널 클라우드	만물인터넷	3D 프린팅
4	비디오	사물인터넷	사물인터넷	하이브리드 클라우드와 서비스 브로커로서의 IT	보편화된 첨단 분석
5	차세대 분석	앱스토어와 마켓플레이스	하이브리드 IT & 클라우드 컴퓨팅	클라우드 / 클라이언트 아키텍처	다양한 정황 정보를 제공하는 콘텍스트 라지 시스템
6	소셜 분석	차세대 분석	전략적 빅데이터	퍼스널 클라우드의 시대	스마트 머신
7	상황인식 컴퓨팅	빅데이터	실용분석	소프트웨어 정의	클라우드 / 클라이언트 컴퓨팅
8	스토리지급 메모리	인메모리 컴퓨팅	인메모리 컴퓨팅	웹스케일 IT	소프트웨어 정의 애플리케이션과 인프라
9	유비쿼터스 컴퓨팅	자전력 서버	통합 생태계	스마트 머신	웹스케일 IT
10	페블릭 기반 컴퓨팅 및 인프라 스트럭처	클라우드 컴퓨팅	엔터프라이즈 앱스토어	3D 프린팅	위험 기반 보안과 자가 방어



주요 기술은 이미 빅데이터와 연관, 융합되어 있음



# 텍스트, 이미지, 동영상, 음성 그리고 숫자 데이터





# 목 차

비즈니스 혁명을 위한 빅데이터 플랫폼 및 활용사례

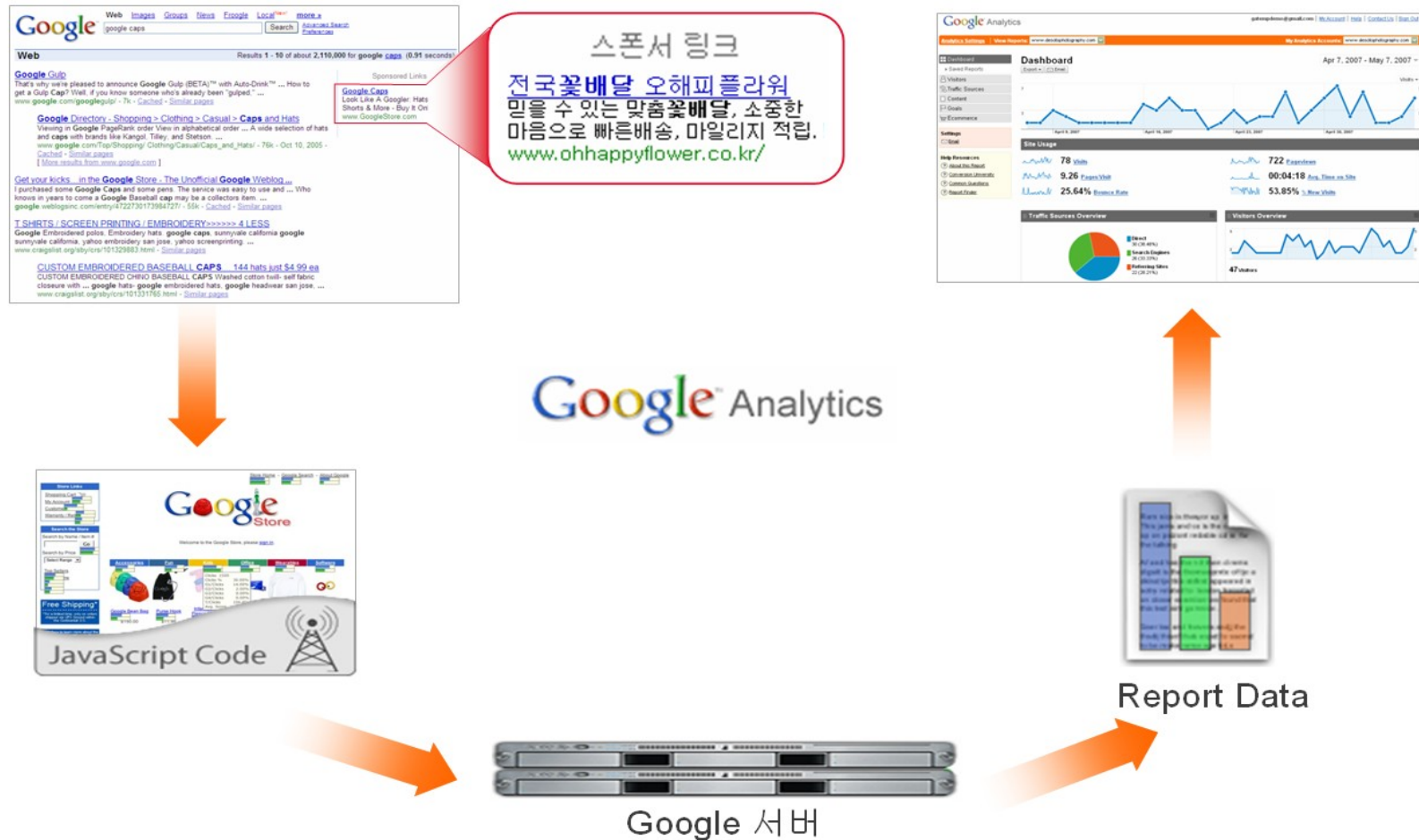
**1 ▶▶ 빅데이터와 비정형데이터**

**2 ▶▶ 빅데이터 플랫폼**

**3 ▶▶ 빅데이터 활용 사례**

# 현재, 시스템 없이 분석하기 - GA

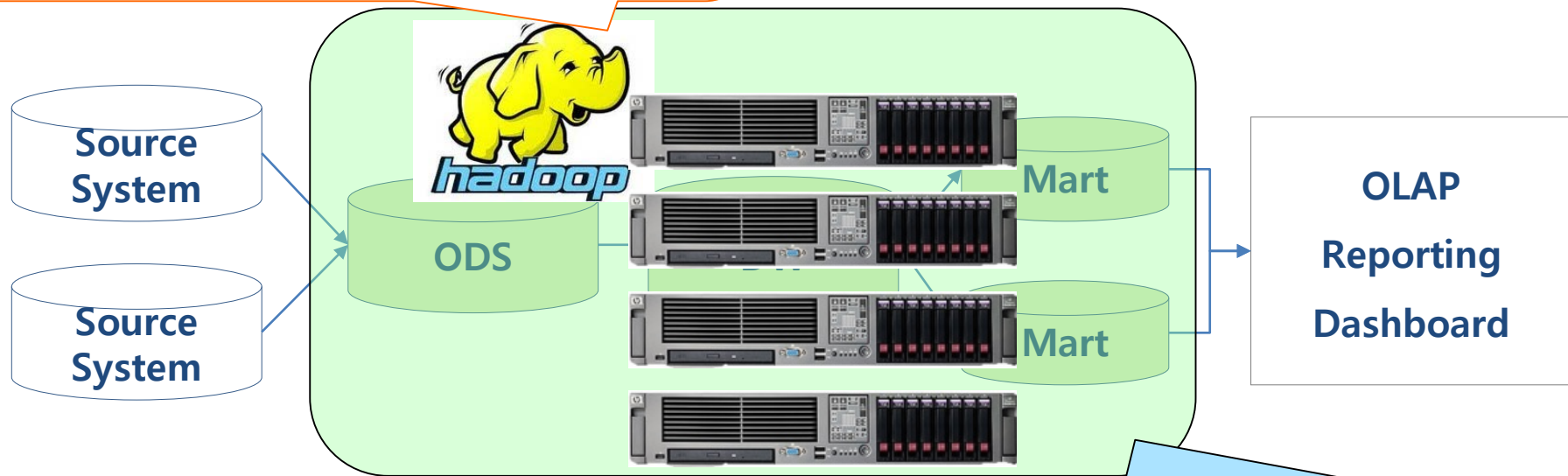
웹 페이지에 GA(Google Analytics) 코드를 추가하기만 하면, 별도 시스템 구축 없이 분석 리포트를 이용할 수 있음



# 하둡과 성능

파일 저장, 대량의 로그, 센서, 텍스트  
데이터 저장에 적합

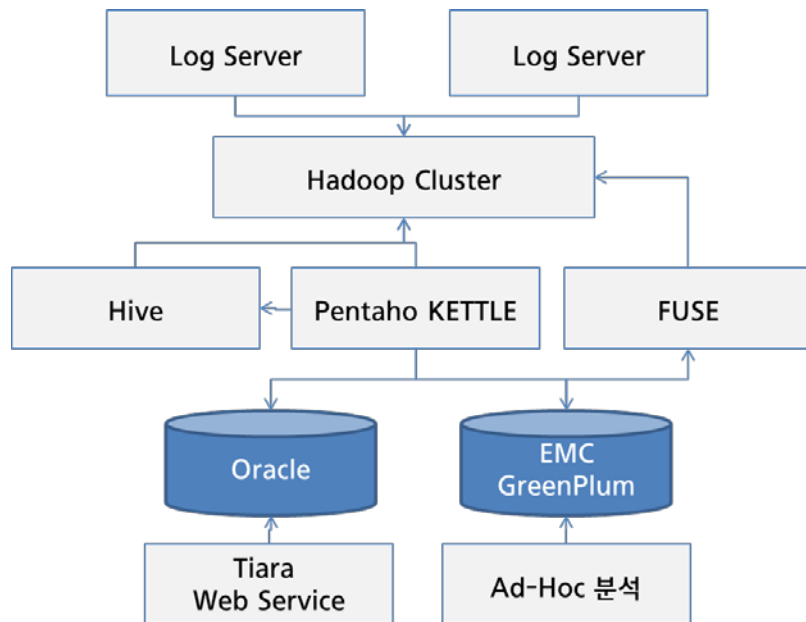
단, 압축 없이 중복 분산 저장(x 3~5)하므로  
저렴하나 많은 디스크 용량을 필요로 함



단순 집계는 매우 빠르다(sum)  
비즈니스 로직을 처리할 경우는 매우 느리다(join, update)  
온라인 분석에 부적합

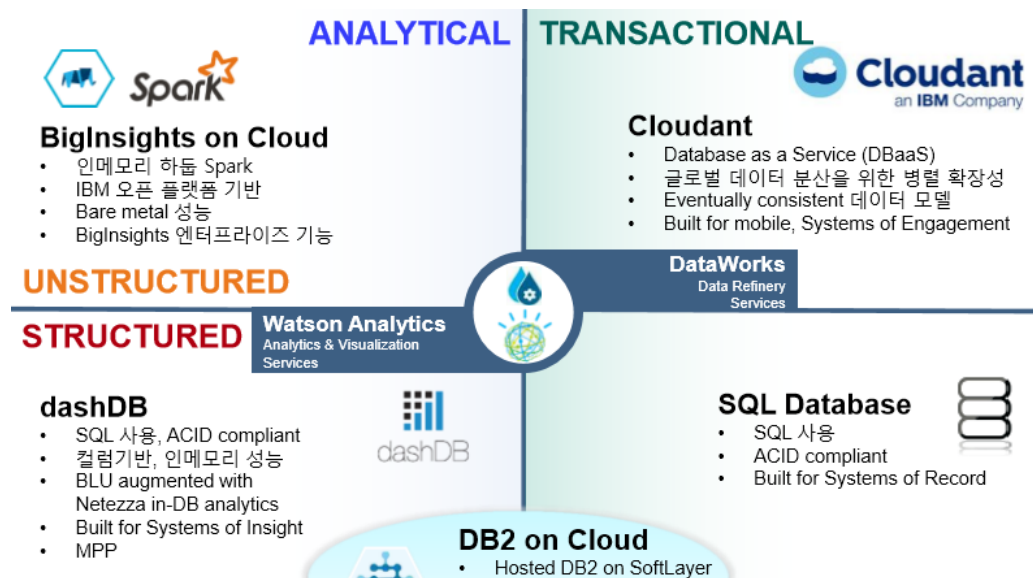
# 플랫폼 관점의 동향

[2010년 전후 웹로그분석 구성도]



여러 기술들이 보완재로 공존

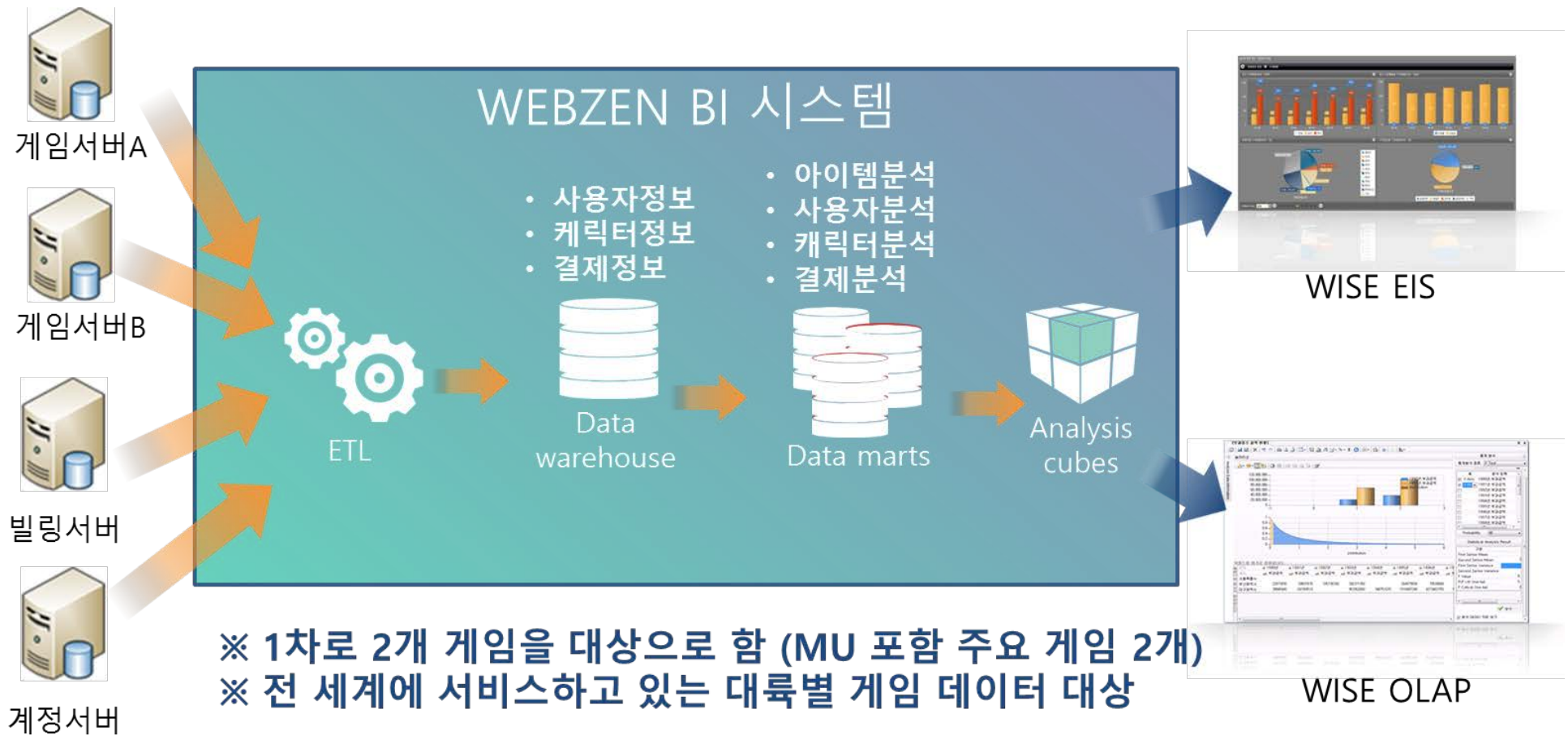
[2015~2017년 IBM 클라우드 구성도]



인공지능과 클라우드

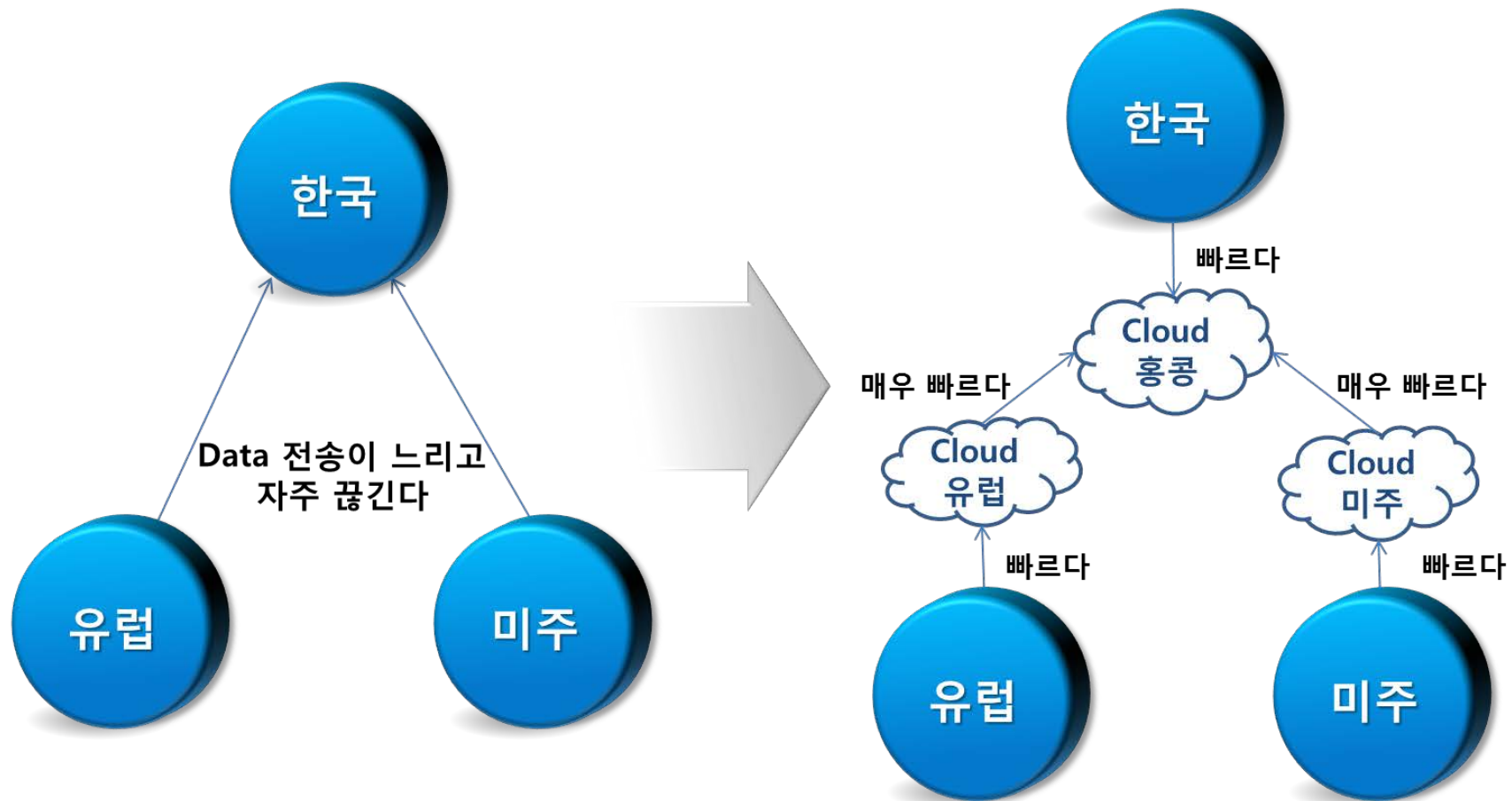
오픈소스와 상용의 혼재

# 게임 데이터 분석





# 게임 데이터 분석 - 대륙간 데이터 통합



# 스마트미디어랩(SMR)

## SMR

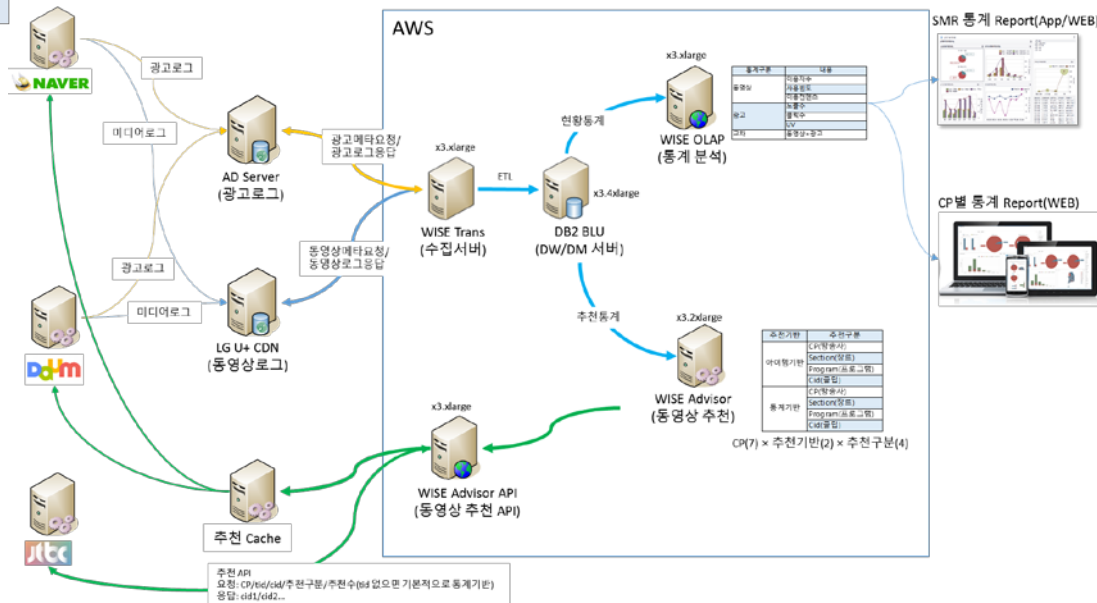
### ### 광고 로그데이터 현황

광고로그	NAVER	GRETECH	DAUM	SMR_MEMBERS	합 계
Row Data (일)	약 100,000,000 건	약 20,000,000 건	약 20,000,000 건	약 5,000,000 건	약 150,000,000 건
Table Max Row Size	10656 byte	10656 byte	10656 byte	10656 byte	
Table Size (일)	50 GB	10 GB	10 GB	2.5 GB	80 GB
Table Size (월)	1550 GB	310 GB	310 GB	80 GB	2500 GB

온라인 미디어 콘텐츠 서비스를 제공하는 SMR은 동영상 및 광고 관련 일 트랜잭션 2억건의 빅데이터 처리의 추천 및 통계분석 처리를 위해 클라우드 인프라 환경에서 WISE Advisor BLU™를 적용하여 업무 요구사항을 충족시켰으며, 빅데이터 플랫폼과 연계를 통해 실시간 동영상 로그 처리와 사용자 맞춤형 추천 서비스를 제공하고 있다.

## 아키텍처 구성

### ➤ 클라우드 환경에서 WISE Advisor BLU적용 사례

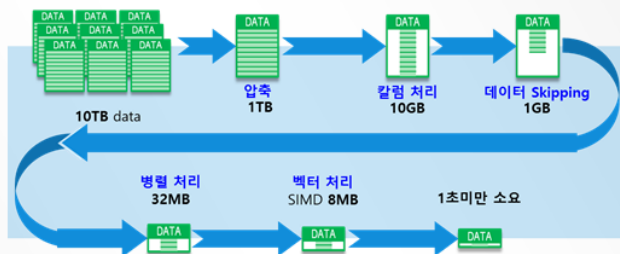


# 빅데이터 분석 플랫폼

하둡, 인-메모리, 컬럼형 DB로 **빅데이터분석 환경**을 제공

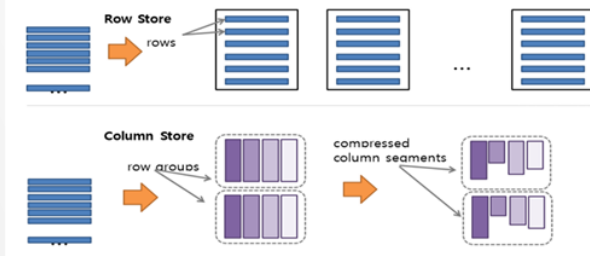
## 인-메모리, 컬럼형 DB

- 인-메모리, 컬럼형 DB 적용하여 획기적으로 분석성능을 개선



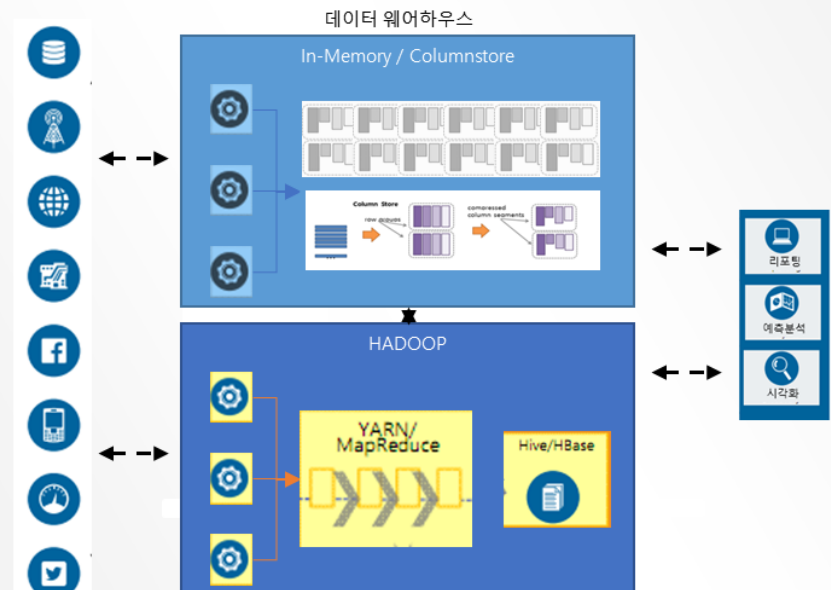
- 시스템 사양 : 32 cores / 1TB memory
- 데이터 크기 : 100 컬럼을 가진 10TB 테이블 ( 10년 동안 축적된 데이터 )
- 쿼리 : 2010년에 얼마나 많은 "SALES"가 일어 났는가 ?  
 • `SELECT COUNT(*) FROM MYTABLE WHERE YEAR = '2010'`

### Rowstore vs Columnstore

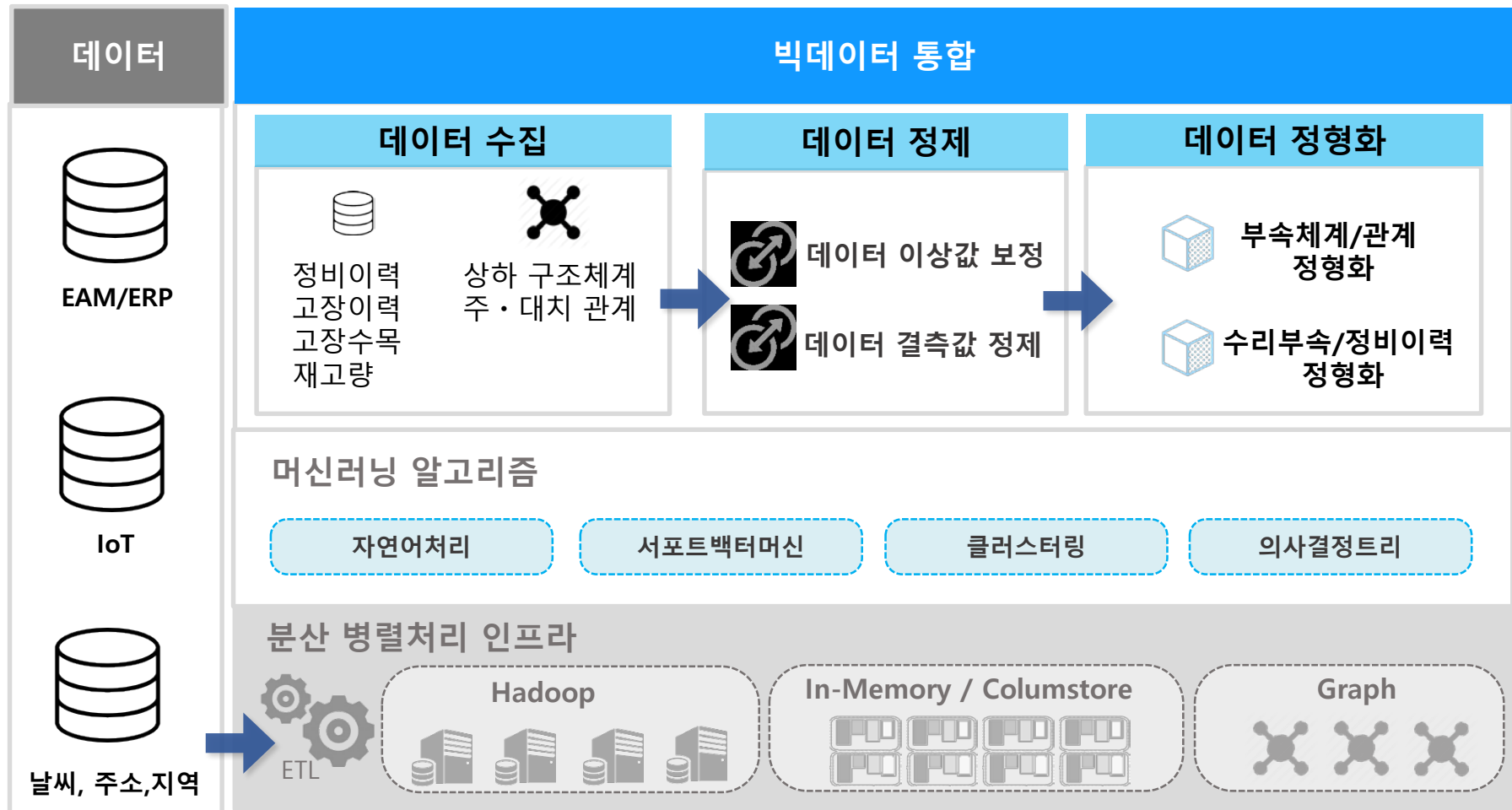


## Hadoop

- 분석과 예측을 위하여 실시간으로 비정형, 정형데이터를 Hadoop 과 애널리틱스 DB에 저장



# 빅데이터 통합 플랫폼



# 목 차

비즈니스 혁명을 위한 빅데이터 플랫폼 및 활용사례

- 1 ▶▶ 빅데이터와 비정형데이터
- 2 ▶▶ 빅데이터 플랫폼
- 3 ▶▶ 빅데이터 활용 사례



# 국방 비정형 데이터 활용

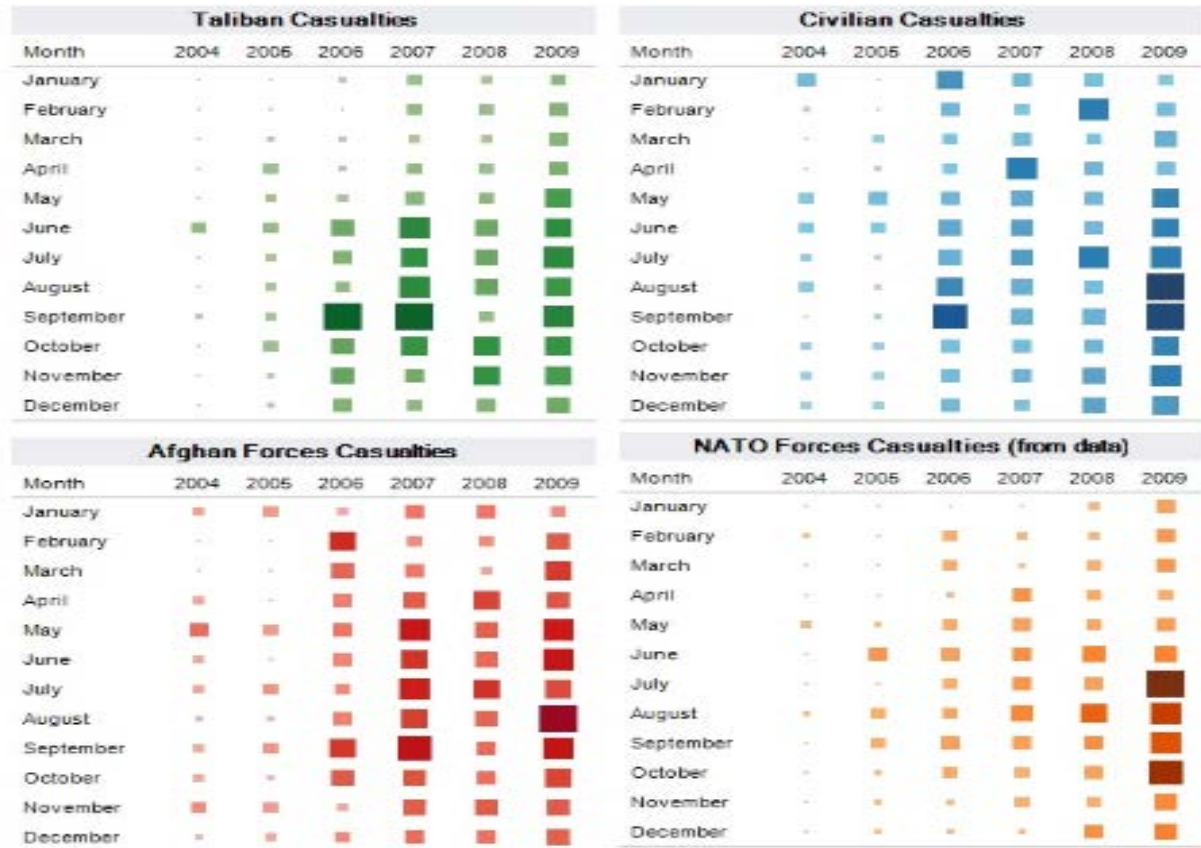
## 국방 - 데이터 분석을 통한 효과적인 전술 정보 제공



드루 콘웨이 (Drew Conway)



위키리크스의 76,900건의 아프가니스탄 전쟁 일지'분석

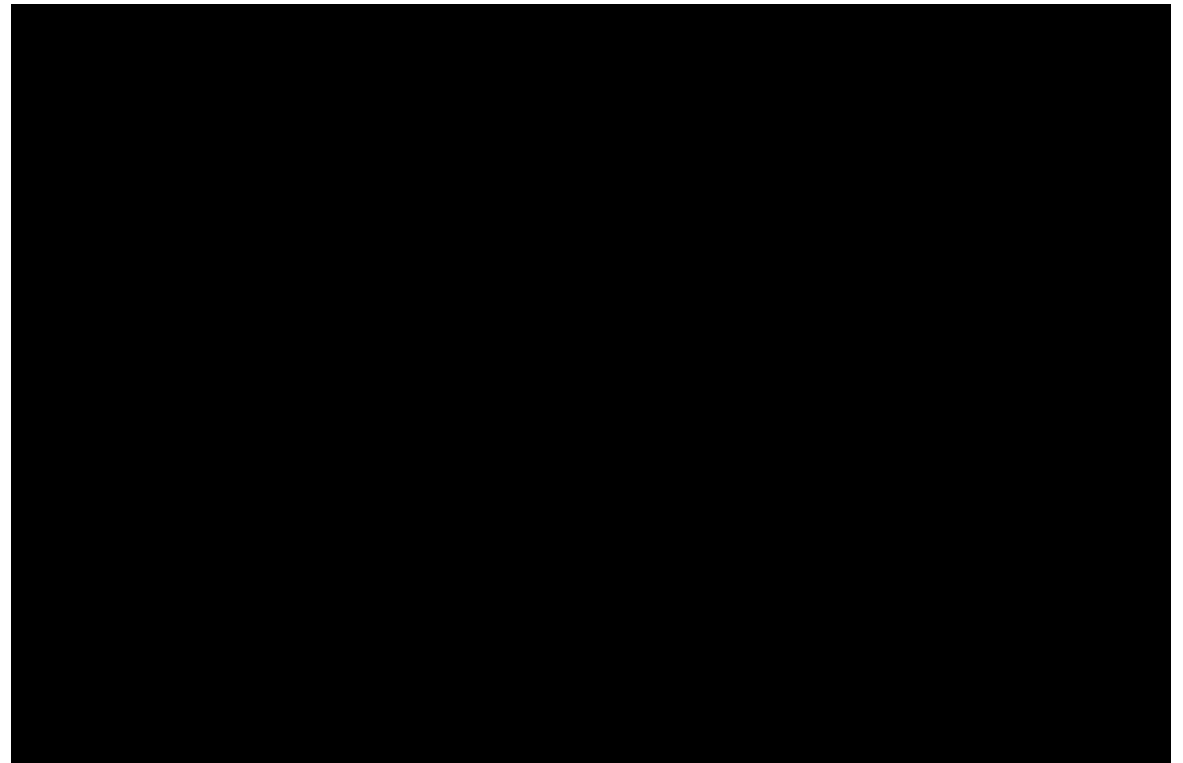
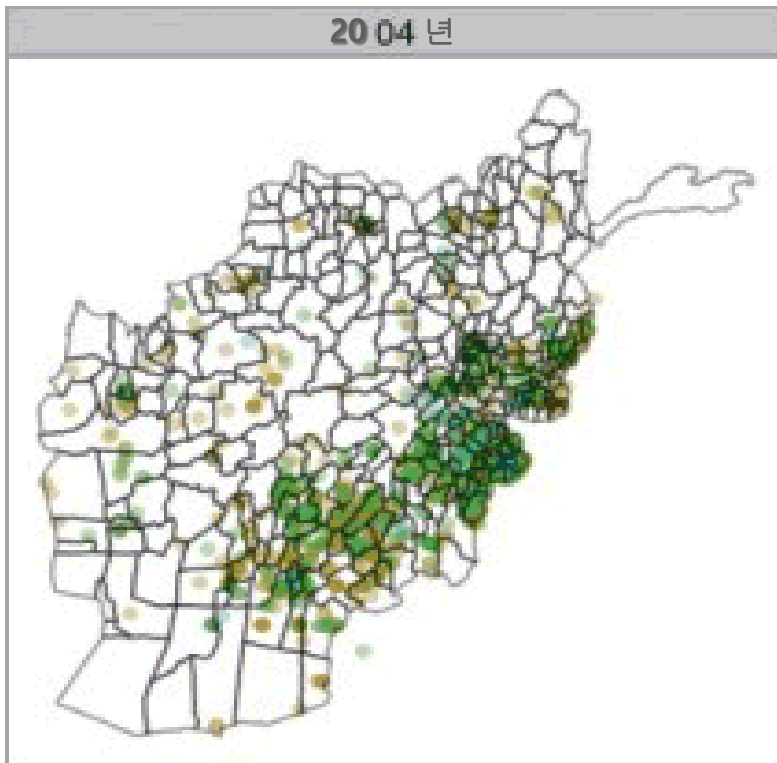


미국과 아프가니스탄 연합군의 병력 활동 동향 패턴 분석, 전쟁 양상 변화 파악

# 국방 비정형 데이터 활용

## 국방 - 데이터 분석을 통한 효과적인 전술 정보 제공

위키리크스 데이터를 분석하여 미국과 아프가니스탄 병력 활동 동향 파악

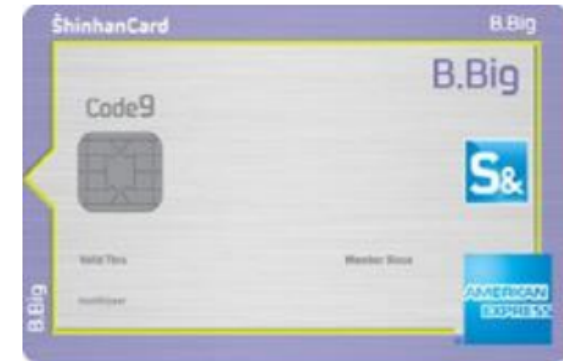


# 국내 신용카드사 사례

## • 국내 신용카드사 빅데이터 분석시스템 도입을 통한 경영혁신 사례

### 카드사 빅데이터 경쟁 치열

- |        |  |
|--------|--|
| 신한카드   | <ul style="list-style-type: none"> <li>업계 최초 빅데이터 솔루션 도입 : 빅데이터 트렌드 연구소 설립</li> <li>- 2,200만 고객 데이터 기반 고객별 소비성향과 유형분석</li> <li>코드나인 매칭 솔루션을 통한 코드나인 카드 300만장 발급</li> <li>'15 빅데이터 경영성과 : 5,700억 카드이용액 창출, 250억 손익개선</li> <li>음성데이터 등 비정형 데이터 분석을 통한 민원저감, 마케팅 성과 향상</li> </ul> |
| 삼성카드   | <ul style="list-style-type: none"> <li>해외 빅데이터 전문가 초빙, BDA(Biz Data Analytics)센터 운영</li> <li>빅데이터 분석체계 '스마트 알고리즘' 활용 7개의 숫자카드 상품 출시</li> <li>상담메모 데이터 등 비정형 데이터 분석 사업 추진 중</li> </ul>  |
| KB국민카드 | <ul style="list-style-type: none"> <li>빅데이터전략센터 확대 개편</li> <li>스마트 오퍼링 시스템 및 음성 상담 문자 전환 서비스 실시</li> <li>20~30대 대상 KB국민 청춘대로카드 출시</li> <li>빅데이터 분석역량 기반 '상품전략컨설팅, 업권분석컨설팅' 서비스 상용화 준비 중</li> </ul>   |
| 우리카드   | <ul style="list-style-type: none"> <li>우량고객군 대상 '그랑블루체크'출시</li> <li>20대 대학생 대상 '썸(SUM)타는 우리 체크카드'출시</li> </ul>   |
| 하나카드   | <ul style="list-style-type: none"> <li>빅데이터 분석을 활용한 O2O 연계서비스로 수익창출 모색</li> <li>스타트업 기업에 빅데이터 분석 서비스 제공 계획</li> <li>날씨 정보 제공 '하나하나호우' 서비스</li> </ul>   |
| 롯데카드   | <ul style="list-style-type: none"> <li>빅데이터 기반 맞춤형 쿠폰 실시간 전송 서비스 '노크'론칭</li> <li>고객맞춤형 카드 추천 서비스 '토픽'출시</li> <li>다음과 '롯데카드 스마트 클러치'개발</li> <li>카드사 최초 원클릭 간편결제 시행</li> </ul>   |



# 비정형 데이터 챗봇 사례

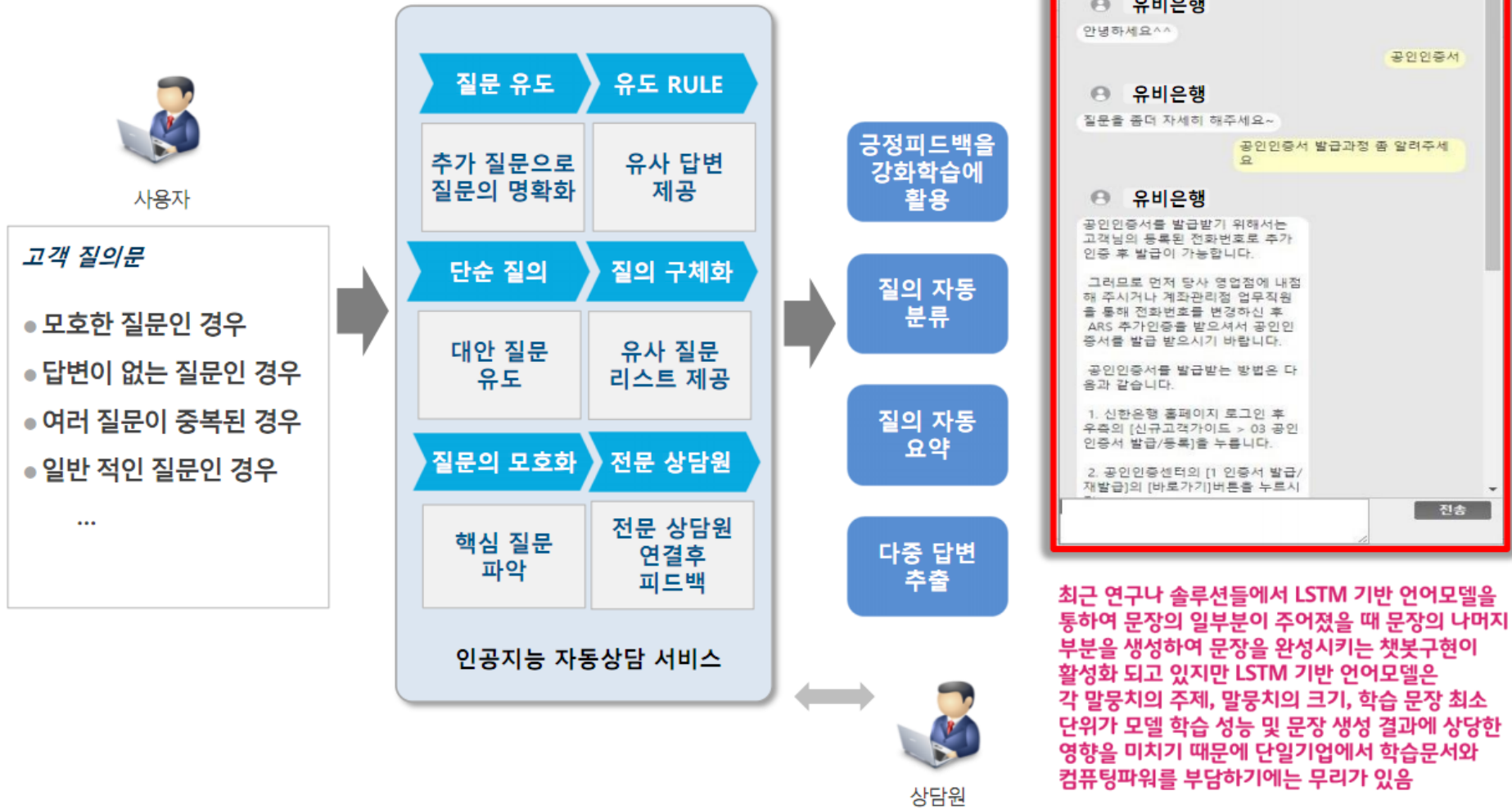
- 챗봇을 통한 고객질의 분석 및 최적의 답변을 제시

금융	• 상품 안내 및 상담, 고객 지원 응대 등
전자상거래	• 쇼핑, 비행기 예약, 숙소 예약, 식당 예약 및 주문, 택시 호출 등
개인 비서	• 헬스케어, 뉴스 피드, 날씨정보, 금융 상담, 일정관리, 길찾기 등
공공 서비스	• 법률 상담, 세금납부, 부동산정보, 구인구직
기업용 메신저	• 정보검색, 파일공유, 데이터 보관, 협업, 고객 관리 등



# 비정형 데이터 챗봇 사례

## 고객 질의 분석 및 최적의 답변 예시



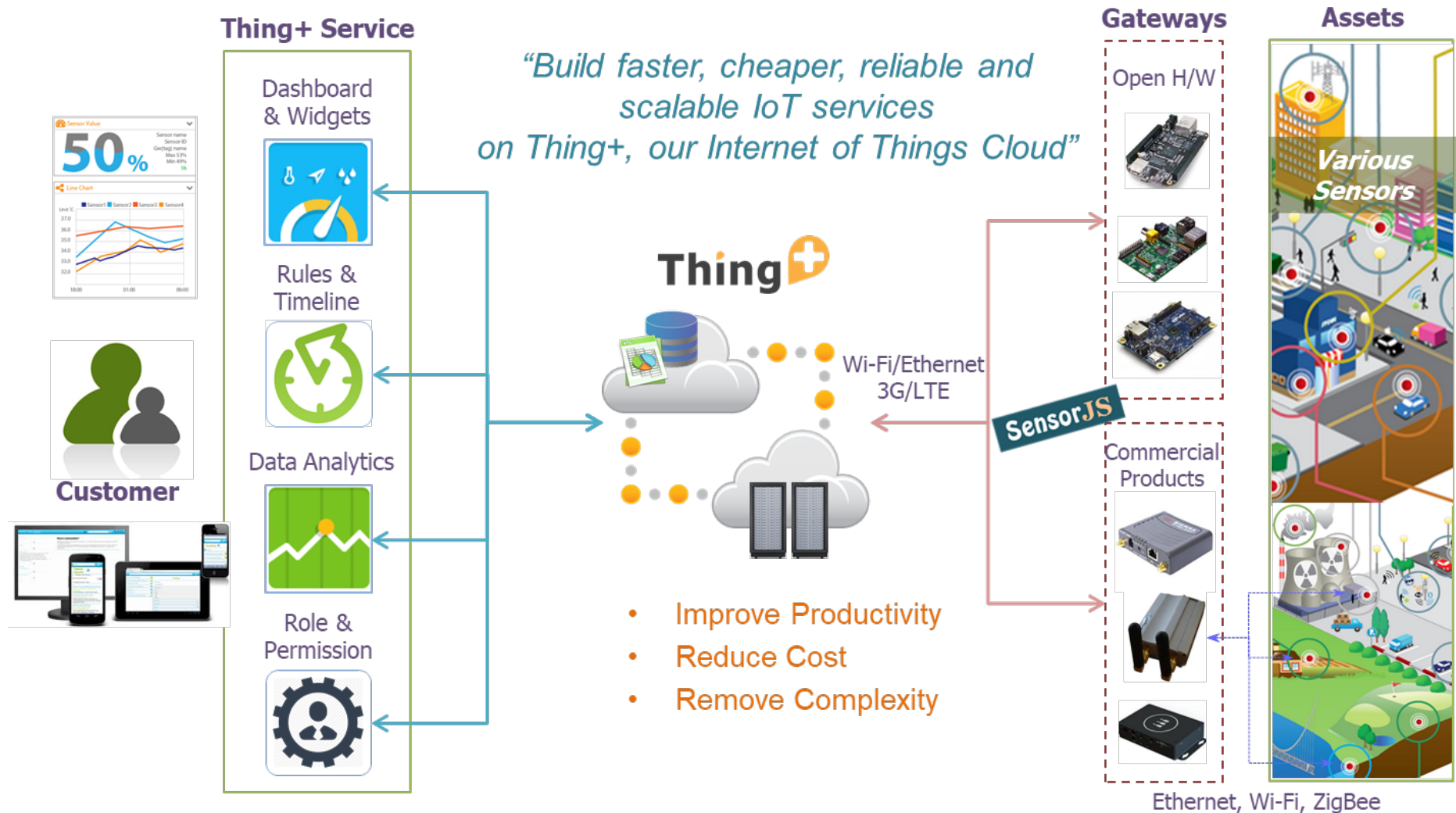
최근 연구나 솔루션들에서 LSTM 기반 언어모델을 통하여 문장의 일부분이 주어졌을 때 문장의 나머지 부분을 생성하여 문장을 완성시키는 챗봇구현이 활성화 되고 있지만 LSTM 기반 언어모델은 각 말뭉치의 주제, 말뭉치의 크기, 학습 문장 최소 단위가 모델 학습 성능 및 문장 생성 결과에 상당한 영향을 미치기 때문에 단일기업에서 학습문서와 컴퓨팅파워를 부담하기에는 무리가 있음

(클라우드가 대세가 되는 이유)



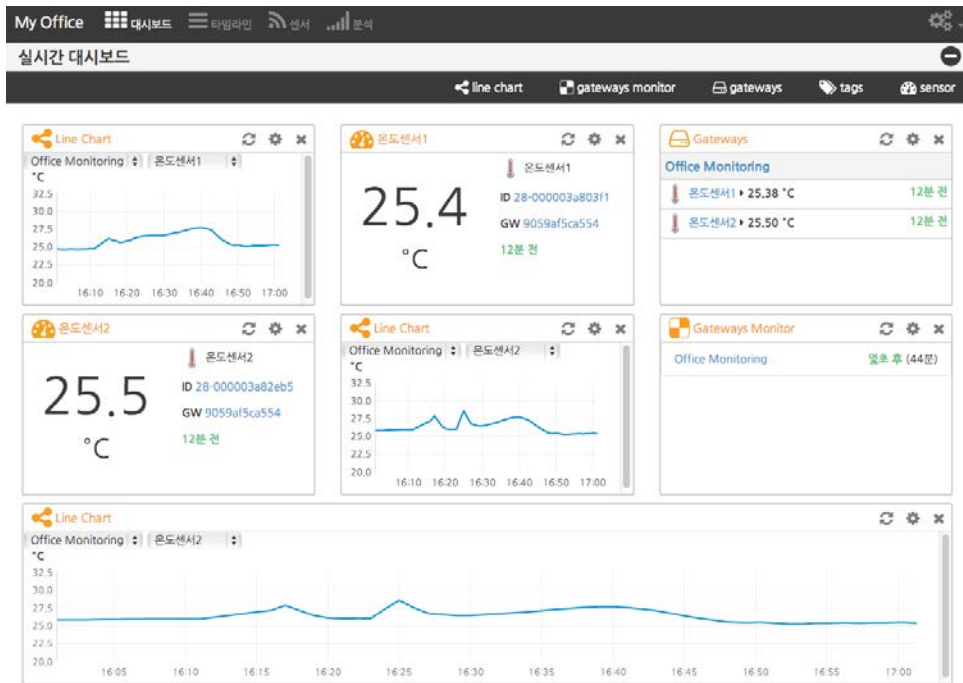
# 센서 데이터 수집 구조 예 - Thing+

## 센서 데이터를 이용한 냉동창고 관리 시스템

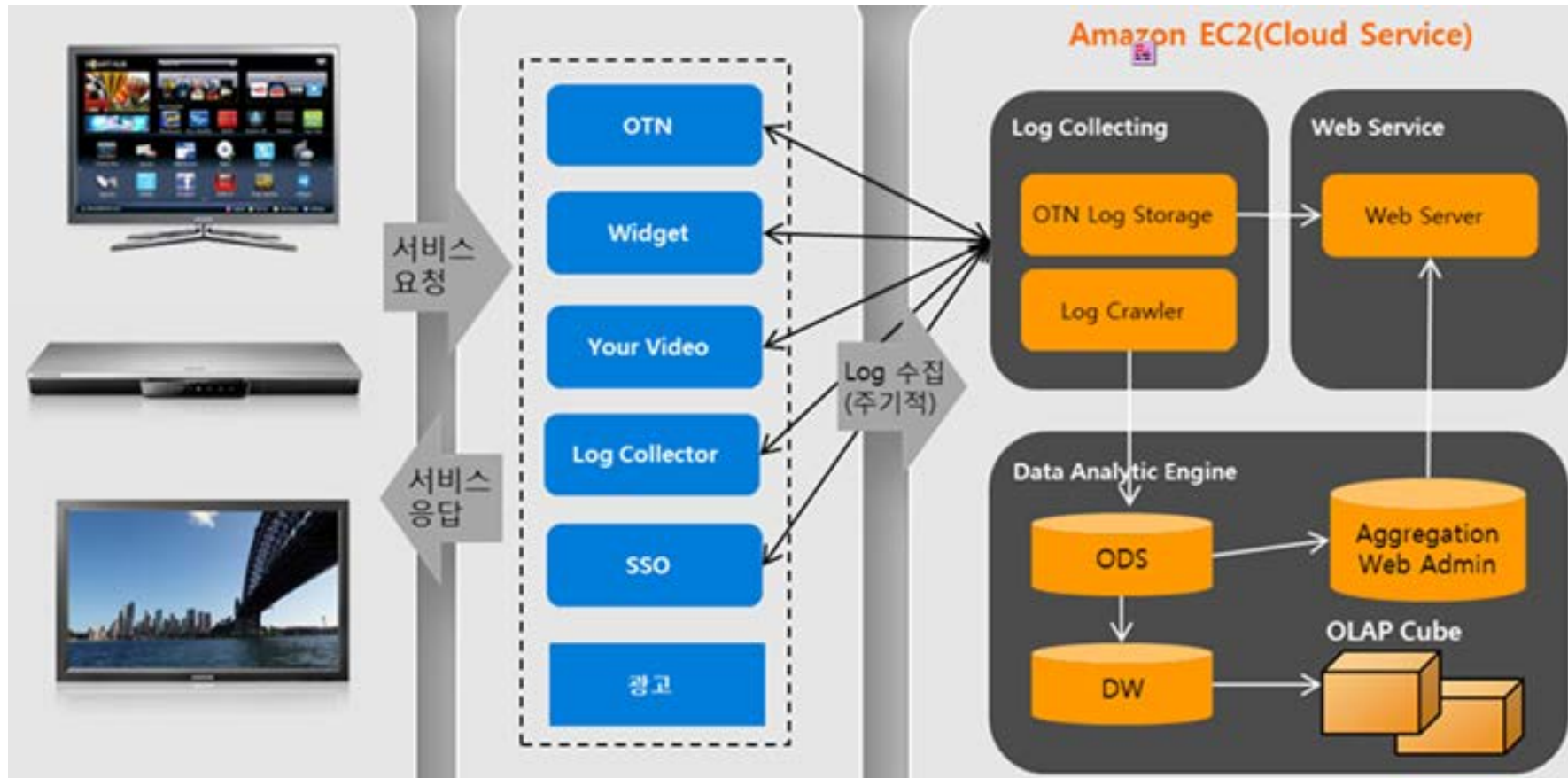




# 센서 데이터 모니터링, 패턴분석 - Thing+



# 로그 데이터 수집 예



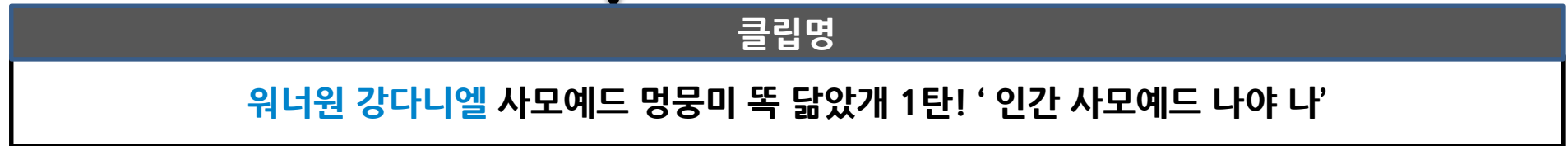
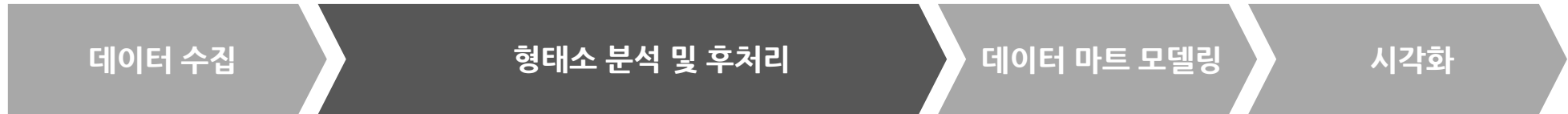
# 미디어 사례 - 키워드 분석

클립 키워드는 해시태그, 검색 키워드, 클립명 정보를 수집함



# 미디어 사례 – 인물사전 형태소 분석

클립 출연진은 형태소 분석과 인물 DB를 활용하여 인물 키워드만 추출함

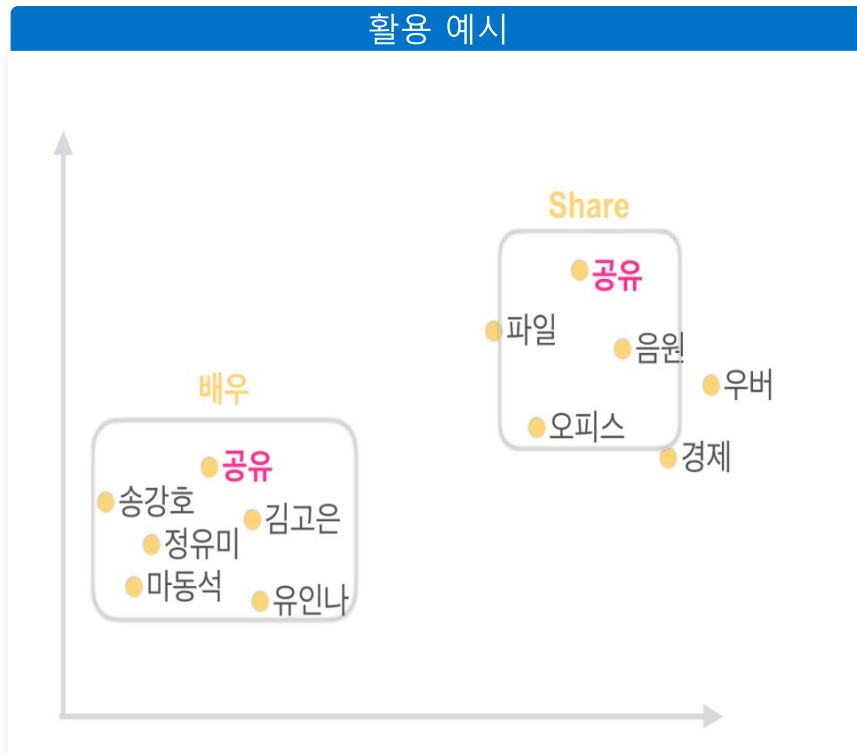


# 메타 관리 시스템 활용



클립, 키워드, 인물 매칭률 증대 방안

활용 예시



공유, 김고은 향한 마음 자각 '첫사랑이었다...'  
tvN 10주년 특별기획 <도깨비>  
▶ 722,388 ♥ 10,412

#김고은 #공유

- 배우 '공유'일 확률 98%
  - 키워드 '공유'일 확률 2%
- ⇒ 배우 '공유' 키워드 타겟팅 연동



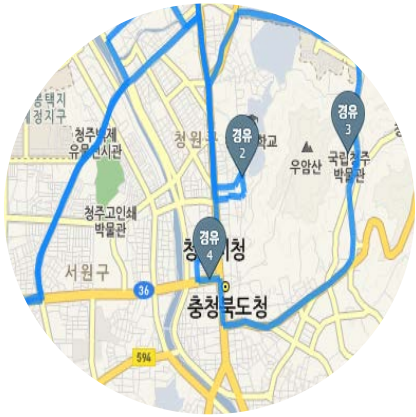
공유 오피스' 임대시장 '열풍'  
디지털조선TV  
▶ 61 ♥ 0

#디지털조선TV #경제뉴스 #금융 #경제 #기업N #임대업 #오피스

- 배우 '공유'일 확률 3%
  - 키워드 '공유'일 확률 90%
- ⇒ 키워드 '공유' 키워드 타겟팅 연동



# 청주시 교통 빅데이터 분석



1.  
청주시  
대중교통  
운행분석



2.  
청주시  
대중교통  
사각지대 분석



3.  
버스 탄력적 배차,  
배차간격 조정



4.  
빵소니범  
찾을 수 있을까?



# 청주시 대중교통 운행분석

청주공항 출발 비행스케줄

상행노선 도착시간	국내선	구분	항공시간
8:00	국내선	출발	7:55
9:30	국내선	출발	8:15
11:15	국내선	출발	8:30
12:45	국제선	출발	10:00
14:30	국내선	출발	11:20
16:00	국제선	출발	11:20
17:45	국제선	출발	11:25
19:15	국내선	출발	11:35
21:00	국내선	출발	11:50
22:30	국제선	출발	12:20
	국내선	출발	13:30
	국제선	출발	13:30
	국내선	출발	15:10
	국내선	출발	17:00
	국내선	출발	17:50
	국내선	출발	18:00
	국내선	출발	18:35
	국내선	출발	18:40
	국내선	출발	19:05
	국내선	출발	20:35
	국제선	출발	21:15
	국제선	출발	21:30
	국제선	출발	22:15

배차시간  
조정 및  
회차 증설

청주공항 도착 비행스케줄

국내선	구분	시간	하행노선 출발시간
국제선	도착	2:30	6:35
국내선	도착	9:00	9:50
국제선	도착	10:15	11:35
국제선	도착	10:25	13:05
국내선	도착	10:35	16:20
국내선	도착	10:50	18:05
국내선	도착	11:10	19:35
국내선	도착	12:35	21:20
국내선	도착	12:55	
국제선	도착	15:00	
국제선	도착	17:00	
국내선	도착	17:10	
국내선	도착	17:25	
국내선	도착	17:50	
국내선	도착	18:00	
국내선	도착	18:30	
국제선	도착	19:00	
국제선	도착	19:30	
국제선	도착	20:15	
국내선	도착	20:30	
국제선	도착	21:25	
국내선	도착	21:30	
국제선	도착	21:35	
국내선	도착	21:40	

회차 증설

배차시간  
조정

버스이용불가

시간축박

오랜시간대기

청주공항<->신탄진  
407번 노선

# 뺑소니범 찾을 수 있을까?

청주시 교통 빅데이터 분석으로  
뺑소니범 찾을 수 있을까



교통데이터수집 장비 DSRC 청주시내 103개 교차로 설치

그러나,

OBU단말기 부착 차량(전체 차량의 약 20%)만 수집

청주시내에 AVI가 많이 설치되었다면  
뺑소니범 찾는데 도움이 되었을텐데 ...



# A마켓

## 추천 입력 데이터



## 추천 입력 데이터 예시

- ▶ 고객정보 : 나이, 성별, 국적 등
- ▶ 조회정보 : 상품의 조회된 정보
- ▶ 선호도 : 고객의 행동 패턴에 따른 제품 선호도 정보
- ▶ 상품정보 : 상품에 관한 정보
- ▶ 유사도 : 상품간의 유사도 정보
- ▶ 상품 추천에 필요한 기타 데이터

## 맞춤 상품 추천 구현 알고리즘

### 클러스터링 알고리즘



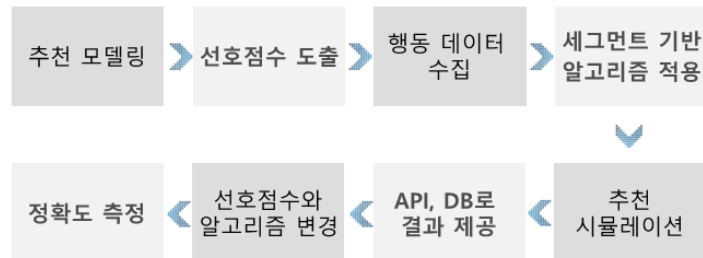
고객 그룹화를 통하여  
정교한 고객  
세그멘테이션 실현

### 협업필터링과 연관규칙



입력된 데이터와 고객  
세그먼트를 활용하여  
개인별 맞춤 상품 추천

## 맞춤 상품 추천 구현 프로세스



## 맞춤형 상품 추천 예시

### 아이템 기반 상품추천



### 통계 기반 상품추천



## 화면 적용

## 맞춤 상품 추천의 유형

추천유형	추천 방법 설명
아이템 기반 상품 추천	<ul style="list-style-type: none"> <li>협업필터링을 이용하여 사용자가 조회한 상품과 유사한 상품 추천</li> <li>연관규칙 알고리즘을 적용하여 동시에 구매되는 확률이 높은 상품 추천</li> </ul>
통계 기반 상품 추천	<ul style="list-style-type: none"> <li>통계적 순위를 기준으로 베스트 상품 추천</li> <li>마케팅에 의해 전략적으로 추천되는 상품 추천</li> </ul>

# 빅데이터와 머신러닝

상품 추천

예측 정비

부당청구 탐지

경마정보 예측

철도 사고  
위험 예측

지능정보형  
인사정책 지원

머신러닝 기반  
데이터 품질

전력이상 상태  
예측

건설 AI 기반  
의사결정 지원

# 감사합니다.

