

# BIG DATA

2015년 빅데이터  
글로벌 사례집

– 분야별 우수사례와 미래부 시범사업을 중심으로 –



미래창조과학부



한국정보화진흥원





# C O N T E N T S

## I 고객관리

1. BoA, 수익성 및 업무효율 제고를 위한 빅데이터 .....	7
2. 허츠, 실시간 VOC 분석으로 고객 만족도 향상 .....	11
3. GS홈쇼핑, 고객 추천 서비스 정교화 .....	15
4. 롯데백화점, 고객 세분화를 통한 타겟 마케팅 .....	19
5. 유통 빅데이터를 통한 중소상인 지원 .....	23
6. 빅데이터 분석 기반 외국인 관광산업 지원 .....	34

## II e-Business

7. Ancestry.com, 온라인 가계도 서비스 제공 .....	49
8. 오비츠, 사용자 특성을 파악하여 맞춤 검색 결과 제공 .....	55
9. NC SOFT, 게임 내 사기 탐지 시스템 구현 .....	59
10. 멜론, 이용자 관심도에 따른 콘텐츠 추천 .....	66

## III 의료

11. UNC헬스케어, 환자의 재입원 비용 절감 .....	73
12. 서울아산병원, 의료연구 편의성 확대 .....	78
13. 맞춤형 유의질병 및 병원정보 제공 .....	82

## IV 제조

14. GE, '지능형 항공 운영' 서비스 .....	89
15. 불보, 운행 정보 활용한 자동차 안전 실현 .....	94
16. 캐터필러, 직원 및 기기 데이터 분석을 통한 제조 생산성 향상 .....	99
17. 한국남동발전, 발전설비 운영효율 극대화 .....	103
18. 자동차 부품기업 공동활용 빅데이터 플랫폼 .....	110

## V 재난·공공

19. 농림수산식품교육문화정보원, 스마트 농정 실현을 위한 플랫폼 구축 .....	125
20. 조류 인플루엔자(AI) 확산 조기대응 .....	133
21. 국도 비탈면 붕괴사고 예측 .....	140



I

# 고객관리





## 1. BoA, 수익성 및 업무효율 제고를 위한 빅데이터

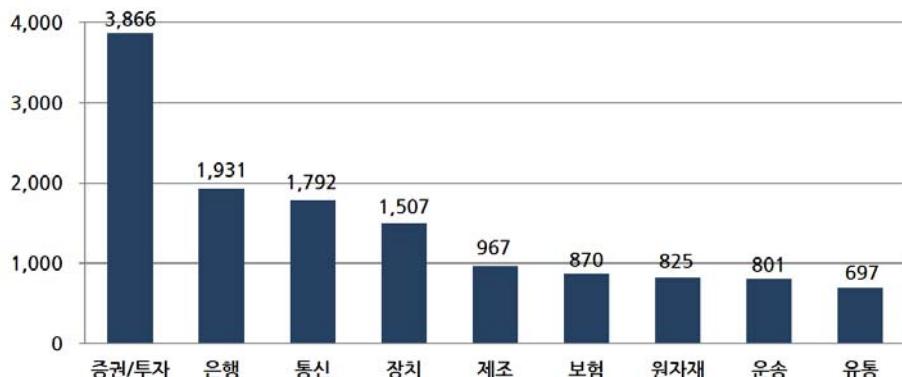


대규모 고객데이터를 기반으로 고객관리, 맞춤형 금융상품 추천 및 신용리스크 조기대응 등 은행 업무 전반의 효율성과 수익성 제고

### 추진 목적 및 배경

- 글로벌 금융기업은 타 산업 대비 높은 데이터 보유량 기록하며 이를 경쟁우위로 활용하기 위한 방안을 강구
  - 증권/투자, 은행, 보험사가 보유한 데이터량은 총 6667TB로 파악되며 전체의 약 50%를 차지
  - 포트폴리오 분석, 트레이딩, 리스크 관리, 마케팅, 보안 등 은행 업무 전반으로 빅데이터 활용 수준을 넓히기 위한 방안 마련 필요

[그림] 미국 산업별 평균 데이터 보유량 (단위: TB)



자료: KB금융지주 경영연구소, 2013

■ 美 뱅크오브아메리카(BoA)는 5천만 건, 약 65PB(petabytes)의 고객 데이터를 보유하고 있으며 이를 분석하여 고객에게 적합한 상품 제안을 하고자 함

- 다양한 채널과 기업 활동을 통해 매우 큰 규모의 고객 데이터를 쌓아왔으나 이를 한꺼번에 분석하는 것은 불가능했기 때문에 표본 분석에만 의존해 옴
- 초기 전체 고객 데이터의 약 1%를 분석에 사용하는 것만으로도 업무개선 아이디어를 얻는데 큰 도움이 되었으며 이러한 경험을 바탕으로 분석 대상 확대 방안 추진

## 추진 내용

■ 빅데이터 활용을 위해 고객에 대한 통합된 접근 방식과 통합된 조직 구조를 마련하고 서비스 설계, 마케팅, 리스크 관리 등 전반적인 활동 수행

■ 빅데이터를 크게 대용량의 거래 데이터, 고객 관련 데이터, 비정형 데이터로 구분한 뒤 정형 데이터인 거래, 고객관련 데이터에 대한 분석을 수행

- 빅데이터 기술을 통해 샘플 데이터가 아닌 고객 데이터 전체에 대한 대규모 데이터 처리와 분석이 가능해짐
- 주요 고객이 어떤 신용카드를 보유하는지, 모기지론을 보유하고 있는지를 파악하고 재 용자가 가능한 지 여부를 결정하기 위해 보유한 트랜잭션과 성향 모델을 활용
- 고객의 쿨센터 이력 정보와 지점 방문 정보 등을 통해 온라인 앱이나 오프라인 판매점을 통해 고객에게 적절한 제안 수행

- 빅데이터 분석가를 확보하고 적극적인 빅데이터 활용을 위해 중앙 분석조직을 마련, 비즈니스 기능 및 단위에 따라 구성된 분석 조직들을 통합하거나 재구성
  - 고객 뱅킹 분석 그룹은 대량 데이터 분석가와 데이터 사이언티스트로 구성되며 업무를 담당하는 임원들과 더욱 밀접하게 관여되어 일을 진행함으로써 효율을 높임
- 빅데이터 활용의 주요 포커스는 고객 상호작용 및 모든 채널을 이용해 고객을 이해하고, 체계적으로 정의된 고객 세그먼트를 대상으로 다양한 금융 상품을 제안
  - 회사가 이미 보유한 트랜잭션 데이터를 새로운 방법으로 활용하여 고객, 판매자, 기업 모두에게 유용한 새로운 서비스를 개발
    - 과거에 고객이 어디서 지출을 했는지에 대한 분석을 기반으로 하며 또한 고객이 분점이나 온라인 채널, 콜센터, 소규모 지점 등 어떤 경로를 통해 유입이 되는지를 이해
    - 이를 바탕으로 고객의 이전 지출 패턴을 도출하여 은행 신용카드 사용자에게 캐시백을 제공하는 'BankAneriDeals'라는 새로운 서비스 제공
    - SNS 등 고객 웹 사용 행적을 분석하여 금융 상품을 고객에게 먼저 제시하는 등 실시간 디지털 마케팅 강화
    - 소셜미디어 분석을 통해 고객 성향을 파악 후 이를 반영하여 자영업자 대상의 자금관리 지원 상품인 'CashPro®Online'과 이의 모바일 버전인 'CashPro®Mobile' 개발
    - 가입자 유치비용은 빅데이터 분석시스템 도입 전에 비해 25% 절감됐고 고객 당 수익성도 12%에서 18%로 증가하는 등 고객유치율과 수익성을 향상
    - 실시간 디지털 마케팅 및 리스크에 대한 조기 경보 체계에도 활용

- 빅데이터 분석시스템을 도입해 신용리스크에 대한 조기경보체계를 강화하였으며 신용관리 및 손실예측 처리시간을 단축
- 대출계좌 40만 건에 대한 신용평가점수를 산출하는 데 걸리던 시간을 3시간에서 단 10분으로 단축
  - 채무 불이행 확률 계산시간을 기존 96시간에서 4시간으로 감소
  - 기존의 임의 처리(Ad Hoc) 분석을 위한 시간을 1/3로 감축

## 효과 및 향후 적용 확대 방안

- 향후 고객들의 SNS 등 비정형데이터를 분석해 고객의 성향과 그날의 기분 등을 파악해 실시간성을 높이는 등 다양한 핀테크 마케팅을 기획
- 빅데이터를 바롯한 다양한 혁신 기술을 적용하여 유연하고 빠른 고객관리와 업무 혁신 지속 수행

## 2. 허츠, 실시간 VOC 분석으로 고객 만족도 향상



세계 각국의 지점을 통해 확보한 방대한 고객 의견을 기반으로  
서비스 개선 및 고객 만족 실현

### 추진 목적 및 배경

- 다국적 자동차 렌탈 서비스 기업인 허츠는 고객 경험의 차별화를 위한 노력 진행
  - 전 세계 146개국에 8,300여개의 지점을 가지고 있는 허츠는 전통적으로 고객만족도 조사를 실시함으로써 고객 유지에 노력
  - 그러나 8,300여개의 지점의 고객만족도 조사가 각각 다루어 졌고, 고객 의견에 대한 통합되고 신뢰성 있는 결론을 얻기는 어려웠음

[그림] 허츠의 다양한 글로벌 고객 데이터 출처



자료: 허츠 발표자료

- 경쟁업체와 차별화된 고객 만족도를 위해 자사가 보유한 빅데이터를 분석하고자 하는 요구 발생

- 매일 수 천 개의 웹 서베이, 이메일, 문자메시지 코멘트를 포함한 엄청난 양의 비정형 자료를 각 지역으로부터 수집
- 가치가 높은 고객의 만족도 정보를 통한 통찰력을 얻기 위해 데이터 통합 및 활용 작업에 착수
  - 기준에 쌓여있던 데이터뿐만 아니라 고객 심리조사를 확대하여 추가적인 데이터 확보에도 노력
  - 최상위 고객을 대상으로는 구조화 되지 않은 자유로운 형식의 피드백 정보를 적극적으로 수집
  - 이러한 데이터를 통합하고 분석하여 각 지역별 적합한 운영 개선 방안을 도출하고 전체적인 서비스 만족도를 개선하기 위한 노력 진행

[그림] 허츠 고객 만족도 측정 및 모니터링 예시



- (1) 이메일/모바일 서베이를 통한 10점 척도의 만족도 조사 실시
- (2) 만족도 점수인 NPS(Net Promoter Score) 도출
- (3) 월간 약 7만 개 이상의 응답 데이터를 축적하여 관리
- (4) 고객 만족도를 실시간으로 트래킹
- (5) 주별 NPS 리뷰를 하여 낮은 스코어를 보유한 지점에 관리 전화 등 정교한 모니터링 진행

자료: 허츠 발표자료

## 추진 내용

### ■ 빅데이터 분석 도구 도입을 통해 의사결정 시간 단축 및 통찰력 확보

- 허츠의 정보수집에 대한 프로세스를 중앙집권화 하기 위해 각 국가와 지역에 고객심리조사에 관한 정보 수집을 강화하고 데이터 별 일관적인 매트릭스를 적용
- 기업 정보에 대한 접근프로세스가 빅데이터 분석도구 도입 이전에 비해 시간이 반으로 줄었으며 이전에는 불가능 했던 다계층 인사이트 확보가 가능해짐

### ■ 통합된 고객 정보를 분석하여 고객의 문제를 실시간으로 파악하고 해결하는데 기여

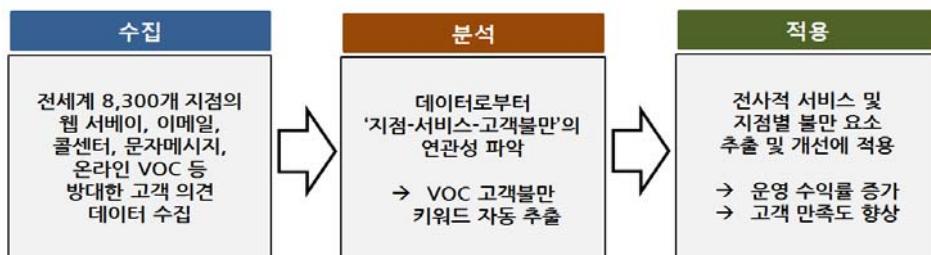
- 다양한 채널에 퍼져 있는 VOC(고객의 소리, Voice of Customer)를 실시간으로 분석해 고객의 요구 사항에 대해 빠르게 대응할 수 있는 시스템을 운영
- 상위 고객으로부터 수집된 구조화되지 않은 피드백 정보를 구조화하고, 지속적으로 그들의 피드백을 분석
- 또한 허츠는 그러한 피드백과 정보에 즉각 응답하여 행동으로 취할 수 있도록 대응 시스템 마련하는 등 분석 결과를 마케팅 및 세일즈를 위한 신속한 의사 결정에 활용

### ■ 전사적 차원의 서비스 개선 뿐 아니라 지역적 특성을 반영한 분석을 통해 실제 수익향상에 긍정적 영향을 줌

- 허츠는 빅데이터 분석을 통하여 필라델피아의 고객 지역이 발생하는 가장 큰 요인이 차량 반납에 걸리는 시간 때문이며, 하루 중 구체적으로 어떤 시간에 이런 지역이 발생하는지 파악
- 이러한 정보를 통해 필라델피아 지점에 고객 집중 시간대의 직원 수를 적절히 조정하고 이슈를 원활하게 해결할 수 있는 지점 매니저를 배치하는 등의 신속한 대응책 실행

- 필라델피아 지점은 이러한 문제를 해결하고 실제 운영 수익률이 증가하는 결과를 보였으며 고객 만족도도 높아짐

[그림] 허츠 빅데이터 적용 프로세스



자료: 허츠 언론보도 재구성

## 효과 및 향후 적용 확대 방안

- 서비스 기업은 다양한 채널로부터 유입되는 고객의 소리를 적극적으로 수용하는 것을 통해 이익 극대화를 기대
  - 고객의 활동 데이터뿐만 아니라 고객이 적극적으로 만들어낸 피드백 데이터나 만족도 데이터를 분석하는 것은 큰 의미가 있음
  - 이를 통해 고객 만족도를 향상시키기 위한 확실한 방안을 확보 가능하며 장기적으로 매출과 직결
- 빅데이터 분석을 통한 통찰력 보유 및 경쟁 우위 확보
  - 시시각각 변화하는 고객의 니즈와 피드백에 대한 즉각적으로 대응이 가능해 진다면 고객으로부터의 신뢰 획득 및 기업 이미지 쇄신에도 긍정적 영향을 줄 것

### 3. GS홈쇼핑, 고객 추천 서비스 정교화



고객 상품 추천 실시간 프로모션 등 하둡 기반의 플랫폼 구축으로  
빅데이터 활용도를 최대화

#### 추진 목적 및 배경

- 고객에게 원하는 걸 쉽게 찾게 해주려는 목적으로 빅데이터에 대한 관심 시작
  - GS홈쇼핑은 홈쇼핑과 인터넷 쇼핑몰인 GS샵 등 TV·인터넷에 걸친 모든 쇼핑 업종을 아우르고 있어 일반 홈쇼핑과 비교해 종류가 다양하며 방대한 고객 데이터 보유
  - 특히 인터넷 쇼핑 비즈니스에서는 데이터를 다루는 문제가 중요하며 기존에는 트랜잭션을 잘 다루는 사이트 관리상의 문제에만 집중했다면 이제는 고객의 행동을 분석하고 추천하는 서비스가 요구되고 있는 상황
  - 이에 회사 보유한 방대한 양의 데이터 분석을 통해 상품 추천 서비스를 시작
- 비용 절감 및 자사 역량 강화를 위해 외부 인력을 최대한 배제하고 자체 기술력으로 하둡 플랫폼 구축
  - 온라인 유통 사업은 판매마진이 크지 않기 때문에 고가의 외산 데이터웨어 하우스 솔루션을 계속 사용하는 것은 부담
  - 또한 전자상거래 기업의 핵심 역량인 상품분석 및 고객 분석을 외부에 의존하는 것은 큰 리스크를 가짐
  - 이에 외부 의존도를 줄이고 자체 기술력으로 상품 추천 플랫폼을 만들기 위한 전략 하에 2012년 초부터 하둡 기반의 오픈소스 시스템을 구축하기 시작

## 추진 내용

### ■ 개별 고객데이터의 면밀한 분석을 기반으로 GS홈쇼핑 상품추천 시스템 개발

- 고객의 클릭이나, 페이지가 넘어갈 때 남는 자취 등 고객 행동을 면밀히 분석하고, 고객이 방문하는 페이지와 페이지 간 연관성을 계산하는 등 다양한 활동을 고객중심으로 연결하는 것을 통해 행동 데이터 추출
- 페이지간 연관성, 유사성을 측정하고 알고리즘화 하며 이러한 작업을 1일 단위로 업데이트
- 고객의 개인 식별은 불가하나 개별 고객의 움직임 패턴을 하나하나 분석하는 방식으로 정확도를 높임

### ■ 하둡 기술자 수급 한계가 가장 큰 어려움. 꾸준한 내부 역량 강화를 위한 노력 진행

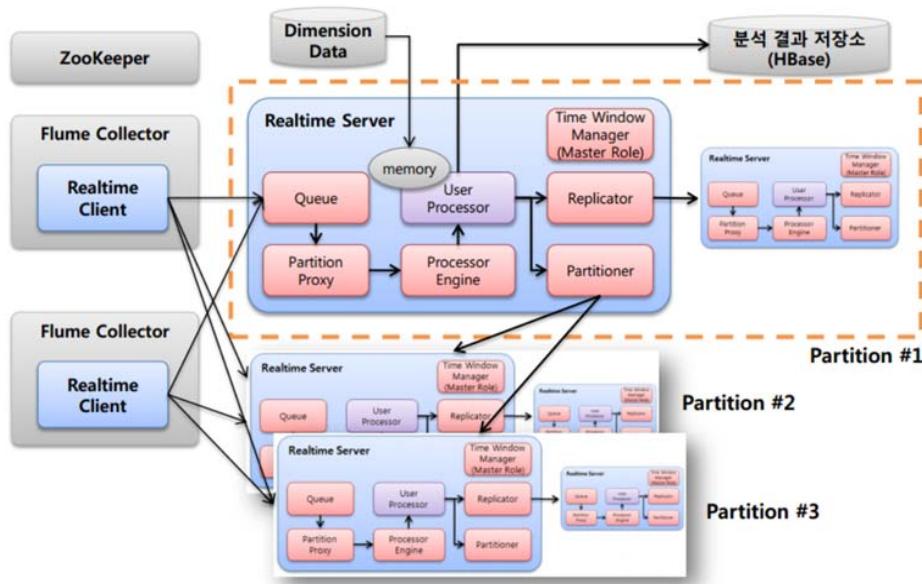
- 빅데이터 도입은 기존 SI 방식으로는 한계가 있으며 기업 내부 인력이 하둡 시스템의 전체 아키텍처를 이해하고 경험을 쌓는 등 자체적인 활용이 가능해야함
- 도입 초기 하둡을 다룰 줄 아는 기술자의 내부 영입의 어려움으로 빅데이터 전문기업인 그루터와의 기술적 협업 진행을 진행하였으며 꾸준히 기술 내재화를 위한 노력 진행
- 장기적으로 외부 솔루션에 의존도를 최대한 낮추기 위해 데이터 사이언티스트 등 빅데이터 기술 및 활용에 전문성이 있는 내부 인력 양성

### ■ 단순히 상품 추천을 위한 빅데이터 적용이 아닌 장기적 관점에서 효과적인 활용을 위한 하둡 플랫폼 구축을 위해 노력

- 과거에는 상품 추천이란 특정 목적에 맞춰 시스템을 운영하는 등 통합적인 활용이 부재하였고, 이에 따라 확장에 대한 유연성도 부족했다는 한계 보유

- 하둡은 플랫폼 위에 필요한 목적과 서비스를 올리는 방식으로 이 시스템을 기반으로 고객들의 다양한 로그 데이터들을 분석하고 고객들에게 맞는 상품을 맞춤형으로 추천해주는 체계 구축
- 초기 목적인 추천 서비스 외 추가적으로 실시간 프로모션 서비스도 하둡 기반으로 구성을 하였으며, 향후 추가적인 목적 발생 시에도 비용 효율적이고 유연한 활용이 가능

[그림] GS홈쇼핑의 분석 플랫폼 아키텍처



- Elastic search / Hadoop / Tajo 등 오픈소스 활용
- 스케일업 전략이 용이하도록 설계
- 통합관리 콘솔을 제공하며, 동시에 협업이 가능하도록 멀티사용자 이용 허가
- 장기적 OLAP 운영을 위한 Tajo 활용
- 충분 데이터와 메타 데이터를 위해 HDFS 저장소 이용
- SQL on Hadoop을 통해 복잡한 MapReduce를 효과적 구현(Tajo)

자료: 그루터, 2013

## 효과 및 향후 적용 확대 방안

- 잘 갖추어진 빅데이터 플랫폼을 활용한 다양한 비즈니스 발굴이 주요 과제
  - 장기적인 관점에서 다양한 목적에 따른 확장이 가능하도록 플랫폼을 구성하였기에 향후 다양한 활용 방향이 기대
- 빅데이터 비즈니스를 지속적으로 창출하기 위한 데이터 사이언티스트 발굴 작업
  - 빅데이터 역량을 강화하기 위한 기술 내재화뿐 아니라 향후 어떤 분야에 빅데이터 기술을 접목할지 찾아내는 ‘데이터 사이언티스트’를 지속적으로 발굴 할 계획

## 4. 롯데백화점, 고객 세분화를 통한 타겟 마케팅



롯데멤버스와 롯데백화점 및 롯데 쇼핑 빅데이터를 활용하여  
고객 특성에 따른 타겟층 선정하고 맞춤형 마케팅 진행

### 추진 목적 및 배경

- 대량으로 축적된 롯데멤버스 회원 데이터를 통한 쇼핑 빅데이터 수집 및 활용 필요
  - 롯데멤버스 카드는 27백만 명의 회원 수를 보유하며 이는 경제인구수의 60% 수준
  - 롯데멤버스 회원의 핫플레이스 이용정보, 엘롯데 고객정보, 외부 가맹점 구매 정보, 롯데 계열사 정보 등이 축적
  - 의미 있는 쇼핑 빅데이터를 모아 빅데이터 CRM(고객관계관리) 개발 및 구축을 위해 활용 계획 착수
  - 장기적으로는 고객 대상의 타깃화된 마케팅을 통한 매출 증대를 기대

### 추진 내용

- 고객의 카드 사용 빅데이터를 수집 및 활용하여 기존 CRM 모델을 업그레이드
  - 고객의 롯데카드 연간 소비액 중 외부 구매율을 반영하여 추가 구매 가능성을 추정하는 쇼핑가능지수를 개발하였으며 이를 기반으로 상향판매(Up-selling) 마케팅 시행
  - 고객별 평균구매주기 및 최대구매주기를 반영하여 고객의 이탈 가능성을 추정하는 이탈 경보 모델(R.E.D Alert<sup>1)</sup>) 프로그램을 활용하여 윈백(Win-Back)<sup>2)</sup> 마케팅 시행

- 구매 패턴을 통해 고객을 분석하고 소비자 이슈 트렌드를 반영하여 이슈 고객에게 마케팅을 시행하는 'L-Trend Catch 프로그램' 시행

[그림] 상향판매 마케팅과 윈백 마케팅 기반



자료: 롯데백화점2014. 3

## ■ 고객의 특성 및 구매패턴 등 다양한 데이터 소스를 활용하여 신규 고객관리 (Customer Relationship Management) 모델 개발

- 연령으로 기반으로 하는 마케팅 한계를 극복하고자, 구매자의 구매 패턴을 분석한 후 구매연령을 기반으로 한 틈새 마케팅 시행
- 고객 특성지수(연령, 성별, 생애단계, 주요 구매 상품 등)와 브랜드 특성지수 (브랜드 M/S, 구매 고객수, 연관 구매 상품군, 인당 구매 금액 등)를 반영한 Shopping Spirit 개발
- 고객별 구매 프로세스에서 겪는 어려움을 고객별로 유형화하고 세분화하여 이에 대응 및 극복 전략을 수립하는 쇼핑 장애요소 극복(Shopping Hurdle) 모델 개발

1) Runaway, Emergency, Devotion Alert

2) 현재 운용 중인 경쟁사의 시스템을 자사의 제품군으로 바꿔 넣는 공격적인 마케팅 방법 [네이버 지식백과]

[그림] 고객 구매 프로세스 유형별 장애요소 극복(Hurdle) 전략



자료: 롯데백화점, 2014. 3

### ■ 쇼핑 빅데이터 분석을 통해 더욱 세분화된 마케팅, 빠른 대응이 가능한 실시간 마케팅 시대로 진입

- 쇼핑 빅데이터를 이용한 원스톱 검색기능과 추천 고객군을 설계하는 원클릭 기능을 더하여 더욱 쉬운 타겟팅 가능
- 실시간으로 데이터를 분석하여 캠페인을 설계하며 이에 대한 반응도 실시간으로 확인하여 빠른 타겟팅 실시
- 다양한 고객군별 선호도를 조사하여 세분화된 타겟팅 구현
- 직관적 분석을 위한 시각화 된 분석 보고서를 구현할 수 있는 시스템 도입
- 상위 1%고객의 취미 등 관심 정보를 등록하여 1:1감성 마케팅을 실시하고 고객 이슈 사항을 메모하는 등 점점 커뮤니케이션 강화하는 VIP 시스템 도입

[그림] 롯데백화점 타겟팅 시스템과 분석 시스템



## 효과 및 향후 적용 확대 방안

- 쇼핑업계의 빅데이터 활용은 고객 구매 패턴 및 인구통계학 자료를 통한 고객 세분화, 타깃 마케팅, 상권 분석, VIP 관리, 비회원·비고객 관리 등 다양한 파생을 기대
- 빅데이터 활용을 통해 고객 소통을 기반으로 원하는 것을 제공하는 차별화 마케팅 제공이 가능하며 기업 타깃이 점차 다수에서 세분화 그룹으로, 또 개인으로 세분화 될 것
- 이를 통해 고객은 적절한 시기에 맞춤 정보 취득 가능, 기업은 수익 극대화와 기업 가치 제고 가능

## 5. 유통 빅데이터를 통한 중소상인 지원



대형유통사의 판매정보를 분석하여 지역 유통시장 중소상인을 위한 맞춤형 상품추천과 시즌기온별 데이터 기반 마케팅 정보 제공

### 추진 목적 및 배경

#### ■ 사업 추진의 배경

- 빅데이터를 적극 활용하는 상위 3개의 유통기업의 생산성과 수익률이 경쟁 기업 대비 6%이상 높음 (대한상공회의소 유통마스터플랜 분야별 정책과제, 2012)
- 대형유통 업계에서는 빅데이터 분석을 활용하여 마케팅 및 매장운영 정책에 반영하여 상당한 효과를 누리게 되었으나, 빅데이터의 혜택을 받지 못하는 중소 유통기업은 경쟁에서 소외되는 현상 발생
- 전체 기업의 60%이상을 차지하며 외부 시장 정보를 전혀 활용하고 있지 않은 중소 업체의 경쟁력 강화를 위한 위한 데이터 분석 서비스가 필요



## ■ 사업 추진의 필요성

- 본 사업 주관기관인 대한상공회의소는 유통사로부터 매월 매장별 매출 데이터를 수집할 수 있는 PDS(Pos Data Service)시스템을 운영하고 있으며 2011년 3월부터 유통시장 분석정보 시스템을 구축 운영 중
- 중소 유통 상공인에게 실질적인 도움을 주고자 기존 보유 데이터와 외부 개방 데이터의 매쉬업(Mash-Up) 분석을 통한 본격적인 빅데이터 서비스 추진

## 추진 내용

| 참여기관 | : 대한상공회의소, 한일네트웍스, 리테일테크, 클루닉스, 디노플러스

## | 주요 활용데이터 |

구분	데이터	데이터 규모	보유기관
매출 데이터	POS(Point of Sale) * 전국 약 700여개 점포, 2015 2월 기준	350MB/주	대한상공회의소
매출메타 데이터	점포정보	210MB/주	대한상공회의소
	점포속성	350MB/주	
	상품분류	70MB/주	
	상품속성	560MB/주	
연계 데이터	기상데이터	3MB/건	기상청
	주민등록인구	10MB/건	안전행정부
	공시지가	5MB/건	한국감정원
	용지별 면적	5MB/건	국토해양부
	Social Network Service	1GB/년	대형포털/커뮤니티

## |분석 내용 및 기법|

### 빅데이터 분석 서비스 모델 선정

■ PDS 자문위원 및 중소 유통사 대상 인터뷰, 벤치마킹, 선진사례 분석을 통해 아래와 같이 서비스와 관련된 주요 진행방향을 도출

- 다양한 외부 데이터와 유통 데이터를 메쉬업하여 실질적인 도움이 가능한 분석 결과를 산출
- SNS 데이터를 활용한 빅데이터와 자연어 처리 기술을 적극 활용하여 연관 상품과 관련된 분석 결과를 도출
- 분석 결과는 IT 접근성과 가독성이 떨어지는 중소 상공인을 고려하여 접근이 쉽고 단순한 유저 인터페이스(UI)를 통해 제공

■ 도출된 시사점과 현재 보유한 데이터, 수집 가능한 외부 데이터를 고려하여 다음과 같이 5개의 빅데이터 주요 분석 서비스 시나리오를 선정함

- 주요 시즌 / 이벤트별 상품군 추천지수 산출
- 기온대별 상품군 추천지수 산출
- 지역특성별 상품군 추천지수 산출
- 신상품 라이프 사이클(Life-Cycle) 매출 추이 분석
- SNS 데이터 기반 연관 상품 분석

### 분석 서비스 시나리오별 상세 내용

■ 주요 시즌 / 이벤트별 상품군 추천지수 산출

- 주요 이슈별 시즌 및 이벤트를 정의하고, 매출 데이터와 시즌 정의 데이터를 연관분석하여 시즌 이벤트별 상품군 추천 지수를 산출

### ■ 기온대별 상품군 추천지수 산출

- 회귀분석 기법을 통해 기온대와 매출량과의 상관관계를 분석하여 5도 단위 기온대를 정의 한 후, 정의된 기온대와 매출 데이터를 연관분석하여 기온대별 상품군 추천 지수를 산출

### ■ 지역 특성별 상품군 추천지수 산출

- 지역 부동산 데이터, 인구 데이터를 종합하여 법정 동 단위의 지역 특성을 선정한 후, 상품군별 전체 평균 매출과 지역 특성에 따른 매출을 비교하여 지역 특성에 따른 상품군 단위의 추천 지수를 산출

### ■ 신상품 라이프 사이클 분석

- 신상품의 출시 후 매출 추이를 동일 상품군 내 다른 상품의 매출 추이 패턴과 비교하여 추후 매출 추이를 예측
- 모든 상품에 대해 출시 후 2년간의 매출 추이 데이터를 산출하고, 이를 상품군 단위로 분류한 후 매출 패턴을 일반화한 4개의 클러스터 매출 추이를 산출
- 최종적으로 신상품에 대한 매출 추이를 해당 상품군의 4개 클러스터 매출 추이와 비교·분석한 데이터를 리포트 형식으로 제공

### ■ SNS 데이터 기반 연관 상품 분석

- SNS 데이터를 수집하여 검색 대상 키워드에 대한 연관 키워드를 필터링하고 해당 키워드 중 음식료품과 관련된 데이터를 추출하여 검색 키워드에 대한 연관 상품을 추출 (ex. 맥주의 연관 상품: 불닭, 소세지, 새우깡)
- SNS 크롤러(Crawler)를 통해 수집된 데이터는 자연어 분석 처리 기법에 따라 연관 키워드를 추출하며, 별도로 작성된 유사어 사전 참조에 의해 최종적인 연관 상품 단어를 추출

### ■ 데이터 처리 및 분석 기법

방대한 양의 매출 빅데이터와 정형/비정형 데이터가 혼재된 데이터를 고속으로 수집 및 분석하기 위하여, 데이터 전처리 기술과 다양한 빅데이터 분석 기법들을 동원하여 데이터를 처리 및 분석함

### ■ ETL 처리

- 스크립트 또는 프로그래밍 언어를 통해 개발한 분석도구를 사용하거나, ETL 전용 분석도구를 사용할 수 있으며, 본 사업에서는 클루닉스사의 MDP 솔루션을 사용

### ■ 클러스터 분석

- 신상품의 매출 패턴 분석 시 동종 상품군의 다른 상품 출시 이후 매출 패턴과의 비교를 위해 비교 대상 상품 전체를 클러스터 분석으로 몇 가지 클러스터(군집)으로 분류

### ■ 자연어 처리

- SNS 크롤러를 통해 수집된 텍스트에서 추출 키워드인 음식료품 단어들을 선별하기 위해 사용됨

## 빅데이터 처리·분석기술 개요

- **ETL(Extraction, Transformation, Load)** 처리는 데이터의 수집, 정제, 단순 변환, 최종 storage로의 데이터 적재 과정을 지칭하고, 데이터를 효율적으로 이전하거나 고속 분석을 위해 불필요한 데이터의 필터링 또는 필요한 데이터의 통합 작업 등을 수행하는 것을 의미하며, 정형 타입의 빅데이터 분석을 위한 전처리 과정으로 수행하는 경우가 많음
- **클러스터(Cluster, 군집)** 분석이란 데이터들의 특성을 고려해 데이터 집단을 정의하고 집단을 대표 할 수 있는 대표점을 찾는 데이터 마이닝 기법중 하나임. 클러스터란 비슷한 특성을 지닌 데이터들의 집단으로 클러스터 분석을 통해 같은 클러스터 내에 특성을 정의 할 수 있고 클러스터 간의 차이를 명확히 볼 수 있음. 이러한 클러스터 분석을 통해 새로운 데이터의 특성을 예측하고 인간이 예측하기 힘든 데이터 그룹간의 차이도 쉽게 파악 할 수 있음
- **자연어(Natural Language)란** 사람들이 일상적으로 쓰는 언어를 일컬어 말하며, 빅데이터 분석에서는 정형화되지 않은 텍스트데이터를 정형화하는 과정을 **자연어 처리**라고 정의함. 자연어 처리에는 형태소 분석, 의미 분석, 대화 분석 등의 분석 방법이 존재하는데, 형태소 분석은 의미가 있는 최소의 단위인 형태소 단위로 문장 성분을 구별하여 분석하는 과정을 말하며 이러한 형태소 분석을 통해 글 전체의 핵심단어를 쉽게 파악하기도 하고 단어와 단어간의 연관성을 추측할 수도 있으며, 이때 분석의 정확도는 얼마나 정확하게 형태소를 구분하여 단어를 추출할 수 있는가에 따라 결정됨

## ■ 데이터 분석 과정

각 분석 시나리오 수행을 위한 데이터의 수집, 저장, 분석 등의 단계에서 필요한 데이터 저장소 및 분석 처리 소프트웨어들을 선정하고, 이를 효율적으로 조작 및 관리할 수 있는 Hadoop 기반의 통합 데이터 저장 분석 프레임워크(G-PAS)을 구축함

## ① 수집

- OpenAPI, FTP, Sqoop을 활용해 내·외부 데이터를 수집시스템으로 수집하는 단계
- 데이터별 수집 주기에 따라 수집을 실시하며, 수집 프로세스는 빅데이터 시스템의 워크플로우 도구로 설정 및 관리

## ② 저장

- Hadoop 기반의 분산 파일시스템을 활용하여 데이터를 저장
- Hadoop은 별도의 백업이 필요 없으며, 일부 구성 노드의 장애시에도 서비스 지속이 가능하며, Hadoop 기반의 분산 병렬 처리 S/W 활용이 용이함

## ③ 정제/변환/매핑

- 분석 과정에 불필요한 데이터를 소거하고, 고속 처리를 위해 필요한 데이터를 변환 또는 병합하는 과정을 수행, ETL 솔루션을 활용함

## ④ 절차적 분석

- 빅데이터에 대한 요약, 평균 산출 등의 기본 통계 수치 산출을 수행하며, ETL 솔루션과 Pig tool을 활용

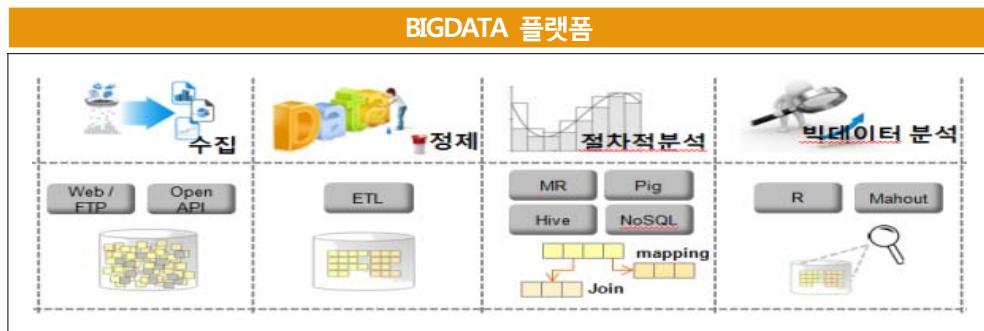
## ⑤ 클러스터 분석 / 자연어 처리

- 신상품 라이프 사이클 분석과 연관 상품 분석에 필요한 분석을 수행하며, 분석 S/W R을 사용함

## ⑥ 결과 업로딩

- 웹서비스나 openAPI를 통해 외부에 제공할 최종적인 분석 결과를 MySQL로 저장하는 과정을 의미하며, Sqoop tool을 사용함

[그림] 대한상공회의소 빅데이터 분석 과정



## 주요 분석 결과 및 활용방안

### | 주요 분석 결과 |

#### ■ 시즌 분석(계절, 명절, 발렌타인데이 등 주요 시즌이나 이벤트 분석)

- 35개 주요 시즌에 대해 114개 상품군 분석결과 쌈장, 김치류 상품군은 여름에 판매량이 가장 높음. 설탕의 경우 초여름(6월 말 경)에 판매량이 가장 높음  
※ 여름철 시원한 음식선호, 여행 증가 등과 설탕절임(잼류 등) 음식 생산 관련된 현상으로 추정

#### ■ 기온별 분석(전체 기온을 5도 단위로 분할하여 매출 분석)

- 35개 아이스크림은 혹한기에 오히려 매출이 증가세로 변함  
※ 혹한기 실내외 온도차에 따른 실내건조 현상으로 인해 소비가 증가하는 것으로 판단
- 차 상품군은 -5도 이하에서 최고 매출을 보이며, 25도 이상에서는 기온상승에 따라 급격한 하락세를 보임

[그림] 주요 분석결과



라이프 사이클, 연관어분석 등 빅데이터 분석 결과를 조회 또는 다운로드 할 수 있는 기능 제공

[그림] 대한상공회의소 유통시장 분석정보 서비스: [bigdata.korcharm.net](http://bigdata.korcharm.net)

## ■ POS 사업자와 협력을 통해 중소매장 POS 단말에 상용 서비스(2015년 1/4분기)

- 데이터 파일 전송 및 API 연계를 통한 중소매장용 POS 시스템으로의 정보제공

[그림] 대한상공회의소 Open API 서비스

## 효과 및 향후 적용 확대 방안

### ■ 사업 의미

- 본 사업은 중소 유통매장 점주들이 원하는 분석 수준과 산출물을 제공
- 최우선 서비스 대상자인 중소 유통매장의 이익 실현, 전체 유통산업 활성화를 통한 경제 활성화에 기여할 수 있는 공공서비스 창출
- 향후 데이터 제공업체/기관의 확대를 통해 전체 유통산업의 세부적인 흐름 까지 분석 활용하여 유통·소매업 경쟁력을 강화

### ■ 활용 및 발전 방안

- 고도화를 통한 서비스 및 리포팅 확장
  - 데이터 추가 확보 및 분석 시나리오 추가 개발을 통해 중소 유통 상공인을 위한 맞춤형 서비스를 지속적으로 확대할 예정
- 데이터 제공을 통한 연계 사업 지원
  - 유통분야 빅데이터 분석 결과 및 빅데이터 분석용 요소데이터들을 필요로 하는 외부 기업이나 기관에 제공하여 2차 사업 활성화 및 빅데이터 서비스 사업에 기여

## 6. 빅데이터 분석 기반 외국인 관광산업 지원



내외국인 관광소비 패턴, 중국인 관광 트렌드를 분석하여 개인 맞춤형 관광정보 제공, 추가 관광지 개발, 관광지 추천 정보 제공

### 추진 목적 및 배경

#### ■ 사업 추진의 배경

- 중국인 관광객 10년전 대비 500% 증가, 외래 관광객 중 가장 높은 비중과 소비규모, 관광 제도 개선으로 인한 자유여행객의 증가 예상
- 외국인 관광객 1000만 시대, 중국인 관광객은 전체 관광객의 33%로 1위
- 1인당 경비 지출이 중국인관광객이 가장 높음 (약256만원, 타 관광객의 40% 이상)
- 중국여유법(旅遊法) 개정으로 단체 여행객의 감소, 개별 여행객 수가 증가하고 있음



## ■ 사업 추진의 필요

- 중국인 관광객의 획일화된 관광지 및 관광패턴, 관광일정으로 인해 재방문률의 정체와 재방문 의향이 낮은 상황
- 국내 입국 관광객의 모바일 인프라를 활용하여 보다 다양한 국내 관광정보의 제공을 통한 관광 만족도를 높일 필요성
- 중국인 관광객의 여행 패턴 및 소비 패턴 분석을 통해 중국인 관광객 대상 정책 활용 및 사업 환경 개선에 활용

## 추진 내용

|참여기관| (주)오픈메이트, 비씨카드(주), 한국관광공사, 나이스평가정보, KT

### |주요 활용 데이터|

구분	데이터	데이터 양	제공 기관
소비 /거래 패턴	내국인 거래 패턴	연 약 24억건	비씨카드
	외국인 거래 패턴	연 약 1,300만건	
	고객 유형 정보 (성, 연령, 주소의 비식별 정보)	약 2,900만건	
관광 권역 및 공간 정보	전국 블록 및 유형 정보	366,999 건	오픈메이트
	주요 상권 영역	1,200 건	
중국어 관광 컨텐츠	중국번체/간체 관광정보 (공통, 이미지, 소개정보, 위치기반 관광정보, 지역기반 관광정보, 숙박, 행사정보 등)	N/A	한국관광공사
유동인구 패턴	내국인 통화 데이터	연 약 1.5억건	KT
	중국인 로밍 데이터	연 약 180만건	
상가/업소 정보	상가/업소 DB	약 300만건	나이스평가정보

## |분석 내용 및 기법|

### ■ 데이터 처리 및 분석 기법

중국인 관광객의 소비패턴, 이동패턴을 파악하기 위해 유동인구 지수를 개발하고  
중국 관광활성화 지역을 추출하기 위한 분석기법 적용

### ■ 유동인구 지수 개발

- 교통개발연구원에서 수행한 「전국교통DB구축사업 교통유발원단위조사 및 기초분석」를 바탕으로 해당 지점 주변의 인구유발시설(백화점, 영화관, 아파트, 지하철역 등)을 찾고 각각의 인구유발시설들이 통행을 유발시키는 영향력을 중첩, 거리에 따른 가중치를 적용하여 도보 가능한 도로에 대해 10m 간격으로 값을 산출, 지수화

### ■ 소비패턴 분석

- 중국인 여행유형별, 여행시즌별, 시간/요일대별 소비패턴 분석, 선호지역 비교 분석. 내국인과의 여행 패턴을 비교하기 위한 내국인 소비패턴 분석도 병행

### ■ 이동패턴 분석

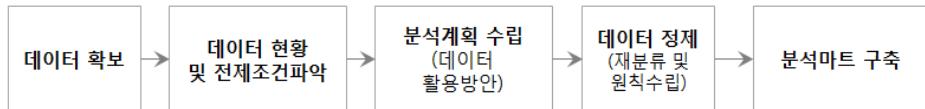
- 카드 거래 데이터 기반으로 소비형태에 의한 이동패턴 분석(여행이동거리, 여행시 먹거리 이동거리 등)

### ■ 관광활성화지역 추출

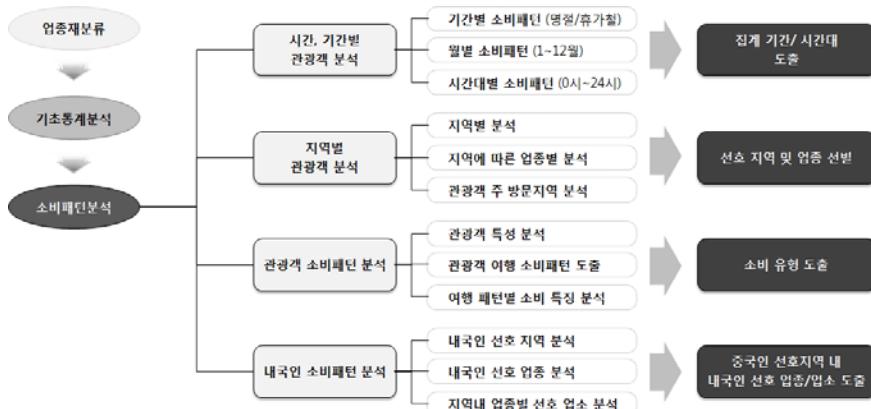
- 서울시 주요여행지역 및 소비밀집도, 통화밀집도 등을 통한 주요 관광권역 분석/추출(6개 주요 권역: 강남/서초, 동대문, 명동/남대문, 이태원, 종로/인사, 홍대/신촌)

## 데이터 처리과정 및 시스템

### • 데이터 마트 구축 단계



### • 데이터 분석과정



### • 데이터 처리를 위한 시스템 구성도



## 데이터 분석 과정

### ■ KT 로밍데이터를 활용한 중국인 유동인구 산출

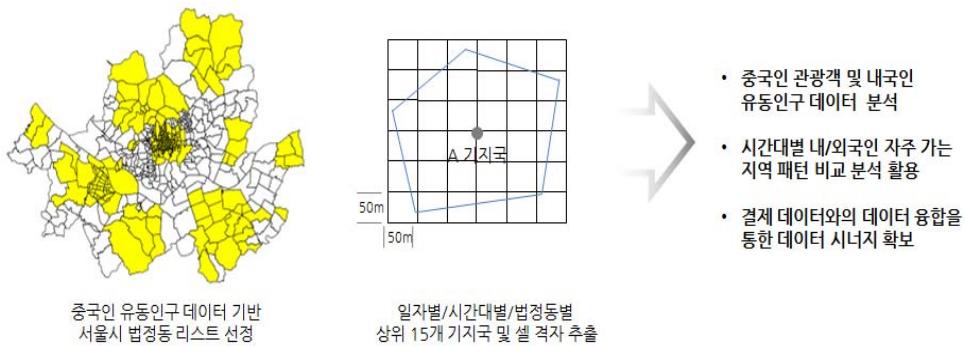
#### ① 유동인구 추출

- 일단위 국제로밍호 테이블에서 서울시 지역에 대한 법정동 단위 유동인구 추출
- BC카드 결제내역 및 동별 시설물 정보 등의 비교 분석을 통한 유의미한 176개 법정동 선정

#### ② 기지국 추출

- 일자별/시간대별/법정동별 유동인구가 높은 상위 15개 기지국 추출
- 통화 발생 기지국 커버리지를 포함하는 50m x 50m 격자셀 단위의 유동 인구 정보 분석

[그림] 유동인구 데이터 분석 과정



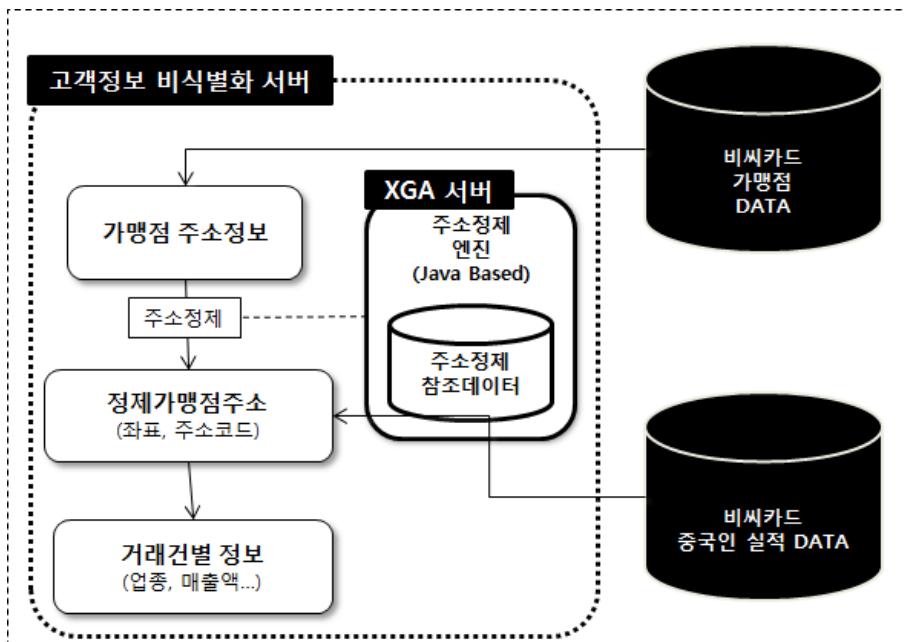
#### ③ 제공 단위 가공 및 분석

- 유의미한 시간구간별 분류 : 00-07시, 07-11시, 11-14시, 14-17시, 17-20시, 20-00시(각 시간 구간별 50m\*50m 셀 단위 집계)

### ■ BC카드 데이터를 활용한 상권별 중국인 매출액 분석

- ① 고객유형정보 분석: 블록 단위(거주지 주소 기준) 성, 연령 고객 유형  
※ 본격적인 정보분석을 위한 개인정보의 비식별화를 비롯한 거래건별 정보 추출
- ② 내국인 거래패턴 분석: 블록 단위 카드 거래의 성별/연령대별, 시간대별/  
요일별 패턴
- ③ 중국인 거래패턴: 블록 단위 은행카드 거래실적 데이터

[그림] 카드정보 분석방법

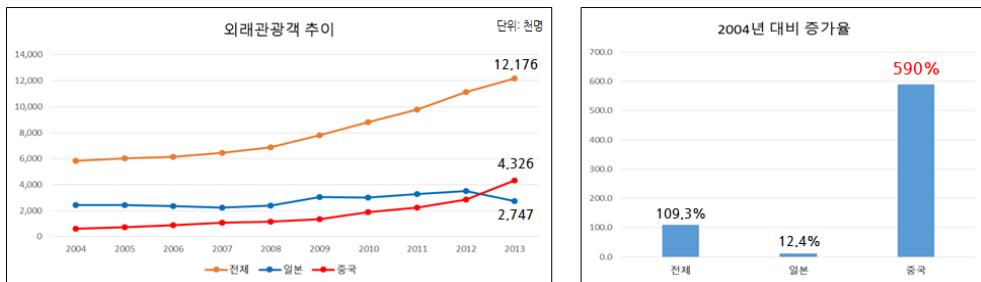


## 주요 분석 결과 및 활용방안

### | 주요 분석 결과 |

#### ■ 중국인 관광패턴 분석

- 외래 관광객은 2008년 이후 매년 10%씩 증가하고 있으며, 특히 중국인 관광객은 500% 이상 증가  
※ 일본 관광객수를 넘어 우리나라 관광산업에서 차지하는 중요도가 크게 상승



- 중국인 관광객 구매횟수는 전년 대비 181% 상승하였고, 구매금액은 전년 대비 120% 상승

#### ■ 중국인 1인당(카드당) 거래 현황

구분	평균 체류기간(일)	평균 방문매장수	평균 구매횟수	평균 구매금액(원)
2013년	2.8	5.9	9.7	2,565,689
2013년 전반기	2.8	5.9	9.5	2,522,406
2014년 전반기	2.9	6.3	10.5	2,235,100
증감	0.1일	0.4개	1회	-317,306

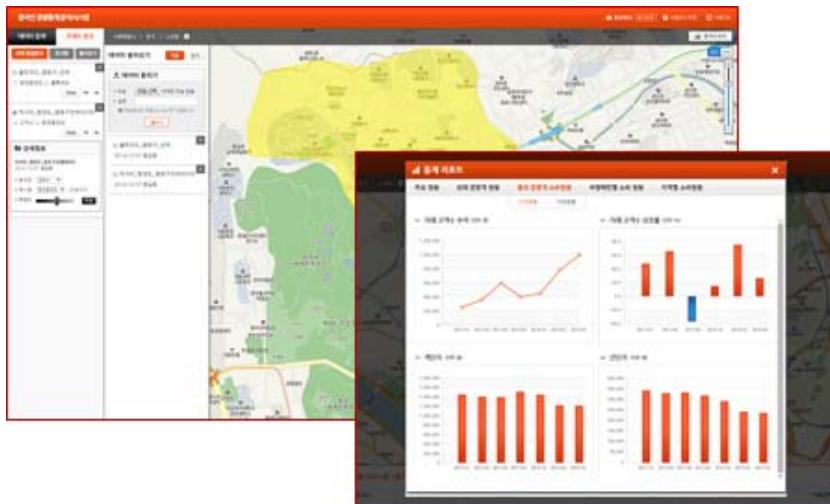
- 명동/동대문 인근 쇼핑타운을 중심으로 밀집도가 높은 것으로 확인되나, 최근에는 강남/서초, 종로권역의 거래증가율이 높게 나타남  
※ 명동/동대문의 경우 백화점/면세점이 월등히 높았으며, 강남/서초는 의복의류
- 중국관광객 평균 구매횟수는 9.7회, 1인당 소비금액은 256만원으로 전체 관광객 평균 183만원 보다 중국관광객이 40% 추가 소비(2013년 기준)
- 관광객은 주로 명동/남대문(2,458,948건)에서 자주 소비하였으며, 종로/인사(288%), 강남/서초(239%)의 거래 건수가 전년대비 크게 증가  
※ 명동/남대문(2,458,948건) → 홍대/신촌(325,006건) → 동대문(322,604건) → 종로/인사(242,882건)  
→ 강남/서초(223,691건)
- 소비금액 역시 명동/남대문(6조 44억 원)이 가장 높게 나타났으며, 종로/인사동(200%)이 전년대비 가장 큰 폭으로 증가
- 명동/남대문(6조 44억 원) → 홍대/신촌(979억 원) → 강남/서초(882억 원)  
→ 종로/인사(450억 원)
- 관광객은 화장품, 의류 등을 주로 구매하고 화장품은 명동/남대문, 종로/인사동, 홍대/신촌에서 의류는 강남/서초, 동대문에서 주로 구매  
※ 업종별 거래건수 : 화장품 판매점 → 여성의류점 → 인삼제품판매점 → 토산품/기념품점 → 할인점
- 강남/서초, 종로/인사동은 한식, 동대문은 갈비·삼겹살, 홍대/신촌에서는 닭갈비를 주로 먹으며, 명동/남대문에서는 커피를 많이 마시는 것으로 나타남

## | 서비스 계획 및 활용방안 |

- 중국인 대상 모바일 앱서비스에 데이터 분석 결과를 활용한 지도기반의 여행지 안내, 여행코스 추천, 인기 가맹점 소개 등 정보 제공
  - \* BC카드의 중국인 대상 앱서비스 '완주인한궈'에 기능 탑재, 한국관광공사의 중국인 앱서비스 '한국 자유여행'에 기능 탑재



■ '유통시장 분관광서비스 관련 기업, 기관, 지자체 등의 의사결정자들을 위한 트렌드 분석 서비스 제공'



## 효과 및 향후 적용 확대 방안

### ■ 지역 소상공인 또는 소규모 지역 단위 현황분석 요구 사용자 지원

- 해당 사업장 주변 또는 원하는 위치의 중국인 관광객의 소비 규모, 트렌드 파악을 용이하게 하여 중국인 관광객 소비에 대한 이해를 도움

### ■ 관광업 관련 종사자 지원

- 중국인 관광객을 대상으로 하는 관광업 관련 종사자에게 중국인 관광객의 유동흐름, 여행패턴별, 휴가시즌별 관광객 특징, 관광지별 유동흐름 및 주변 소비특징 등의 분석을 통해 서울지역 관광패턴과 선호 관광지 등의 정보를 제공함으로써 중국인 관광객의 관광 및 소비 트렌드를 반영한 관광 상품 개발에 활용

### ■ 관광관련 정책수립 지원

- 관광관련 정책수립을 위한 각 지역별 통계정보를 제공함으로써 해당 공공 기관이 관할하고 있는 지역 내 소비규모, 시간대별 유동객, 선호 업종 등을 분석하여 정책 수립 시 활용함으로써 중국인 관광객의 해당 지역 내 관광 편의 및 지역경제 활성화에 활용

<참고>

## 강남스타일 즐기는… 멋 좀 아는 유커들

서동일 dong@donga.com ·김재형 기자 2015-01-21

### 中관광객 서울 소비패턴 빅데이터 분석해보니





동아일보가 단독 입수한 미래창조과학부·한국정보화진흥원(NIA)의 '빅 데이터 기반의 외국인 관광 산업 지원 시범사업' 결과 보고서에 따른 해석이다. NIA는 지난해 1월부터 올 6월까지 18개월간 중국의 은련(銀聯)카드 거래 명세를 기반으로 유커들이 서울 어느 지역에서 어떤 물건을 사고, 어떤 음식을 먹는지 분석했다. 은련카드는 중국인 90% 이상이 사용하는 것으로 알려진 신용카드다. (중략)

유커들의 관광 행태 중 권역별 차이가 뚜렷한 것은 음식이었다. 강남에서 가장 많이 먹는 음식은 양식(43%)이었다. 강남역 인근 파스타·바비큐 전문점 'Big PLATO' 직원 이민영 씨(26·여)는 "중국인들은 주로 모바일 앱을 통해 우리 가게 위치를 확인한 다음 찾아온다"며 "최근에는 유커들을 대상으로 한 음식점 추천 앱을 만드는 관계자들도 '식당 정보 업데이트를 하고 싶다'며 찾아오고 있다"고 말했다. 강남권에서 눈에 띄는 것은 '계장 전문 음식점'(4.3%)이 5위를 차지했다는 것. 강남의 일부 간장계장 전문점은 중국 홍콩 대만 등 동아시아권 관광객들에게 관광 명소로 알려져 있다.

신촌·홍대앞을 찾은 유커들은 닭요리(17.5%)를 가장 많이 먹었고 한식(9.8%) 전문점을 많이 찾았다. 반면 유커들의 '메카' 명동과 동대문에선 카페(39.8%)에서 쓴 돈이 가장 많았다. (중략)

동아일보 기사원문: <http://news.donga.com/3/02/20141120/68005542/1>



II

# e-Business





## 7. Ancestry.com, 온라인 가계도 서비스



생년월일, 출생·사망 기록 등 역사적인 기록 자료 및 유전자 정보 등 다양한 비정형 데이터들의 연관성을 분석하고 검색 행적을 기록하여 조상 정보 찾기 서비스를 제공

### 추진 목적 및 배경

- 방대한 인구 데이터를 저장하고 공유할 수 있는 기술의 발전과 이를 활용하려는 비즈니스 발생
  - 미국, 호주, 캐나다 등 이민자들로 형성된 나라에서 자신의 뿌리 찾기에 대한 관심 증가
  - 고성능 문서스캔과 이를 판독 및 분석할 수 있는 다양한 기술이 발전되고, 미국에서는 이민입국심사 서류, 재판기록 등 각종 공문서가 일반에 공개가 가능해짐
  - 이에 Ancestry.com은 미국과 캐나다, 유럽, 호주에서 고객의 뿌리를 찾아주는 비즈니스 모델을 개발하여 사업을 전개하였으며 현재 200만 명 이상의 회원을 보유
- 가족(조상) 히스토리 정보를 구축하고 개인별 맞춤 서비스를 제공하기 위한 데이터 스토어 구축작업 시작
  - 생년월일, 출생 및 사망 기록, 센서스, 군적기록, 이민 기록, 전쟁 등의 역사 기록 까지 총 120억 건의 방대한 데이터베이스 축적
  - 고문서 연구기관이나 지방자치단체의 허가를 받아 자료 저작권을 직접 매입 하여 활용하는 등 데이터 출처를 다양화
  - 축적된 데이터는 4페타바이트급으로 10페타바이트(10,000,000GB) 데이터 저 장소에 보관 중

## 추진 내용

[그림] 고객 데이터 분석 사례



자료: Ancestry.com

### ■ DNA 시퀀싱을 통한 조상 찾기 서비스 제공

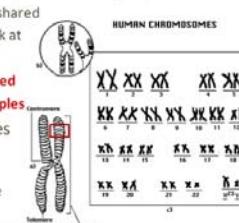
- 회사는 고객이 보내온 태액이 담긴 튜브에 대해 개별적으로 분자테스트를 하여 유전적 데이터를 축적하고 DNA 시퀀싱<sup>3)</sup> 정보분석을 시행
- 모든 AncestryDNA 고객의 경우, 70만 개의 SNP(DNA에 있는 개인 식별이 가능한 변수 영역)들이 측정되며, 이 정보는 회사가 보유한 DNA정보가 있는 모든 가입자들과 비교되어 측정  
※ \$99로 전체 DNA분석의 1/10 가격으로 유전자 분석을 수행
- 사용자의 민족을 예측하고 데이터베이스 내의 친척들을 확인하기 위한 컴퓨터적인 분석도 함께 수행

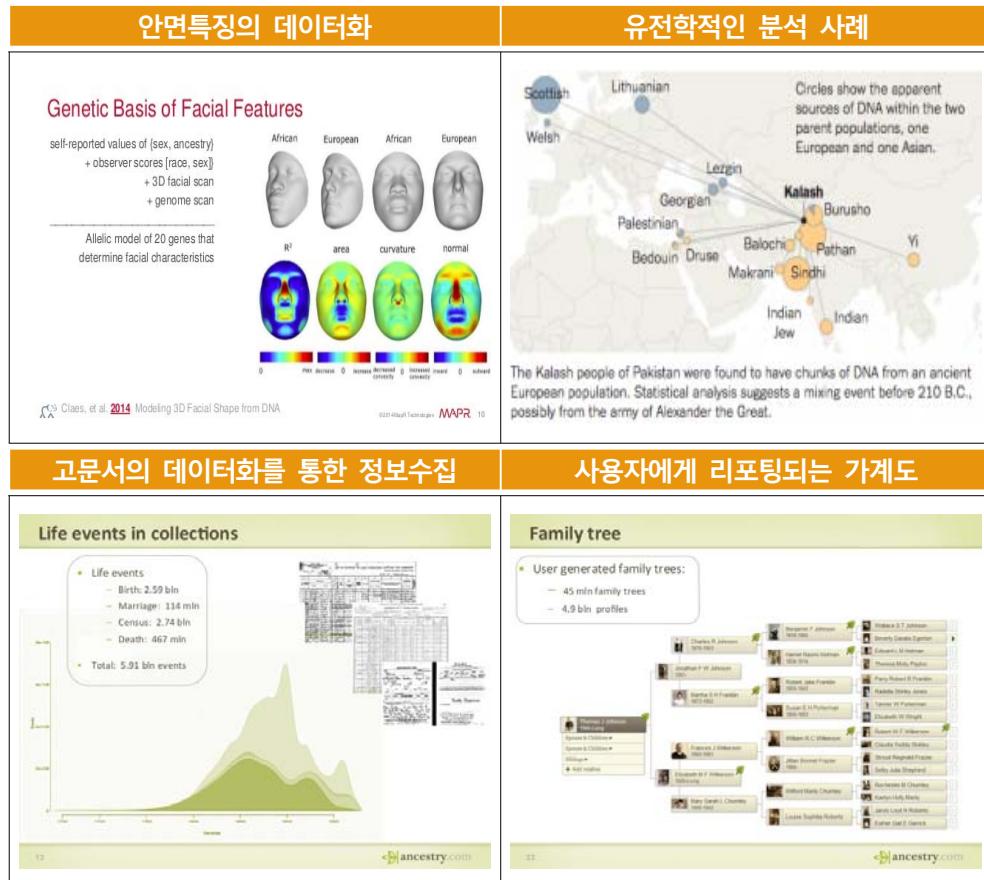
3) DNA시퀀싱 : 생화학적 방법을 사용해 디엔에이(DNA)의 염기서열을 결정하는 과정 [네이버 지식백과]

## ■ 고문서와 온라인상의 조상관련 정보 서비스 제공

- 크롤링(Crawling)을 통해 수집된 디지털 데이터, 스캔된 행정 문서를 활용하여 작성된 온라인데이터베이스에 사용자들이 제공한 데이터를 결합하여 조상과 관련된 콘텐츠 서비스를 제공
- 빅데이터 기술을 활용하여 기록간의 연결고리나 검색 관련 알고리즘 같은 규칙과 절차를 확립. Ancestry.com의 검색 결과는 전략적으로 연결된 기록들과 과거 검색 행동을 기반으로 이루어짐
- 연관어 검색 등 검색 기술도 다양하게 반영하여 검색 정확도를 높였으며 최근에는 방대한 양의 사용자 정보도 추가로 유입하는 등 정확도를 높이기 위한 작업을 지속적으로 진행

[그림] Ancestry.com의 주요내용

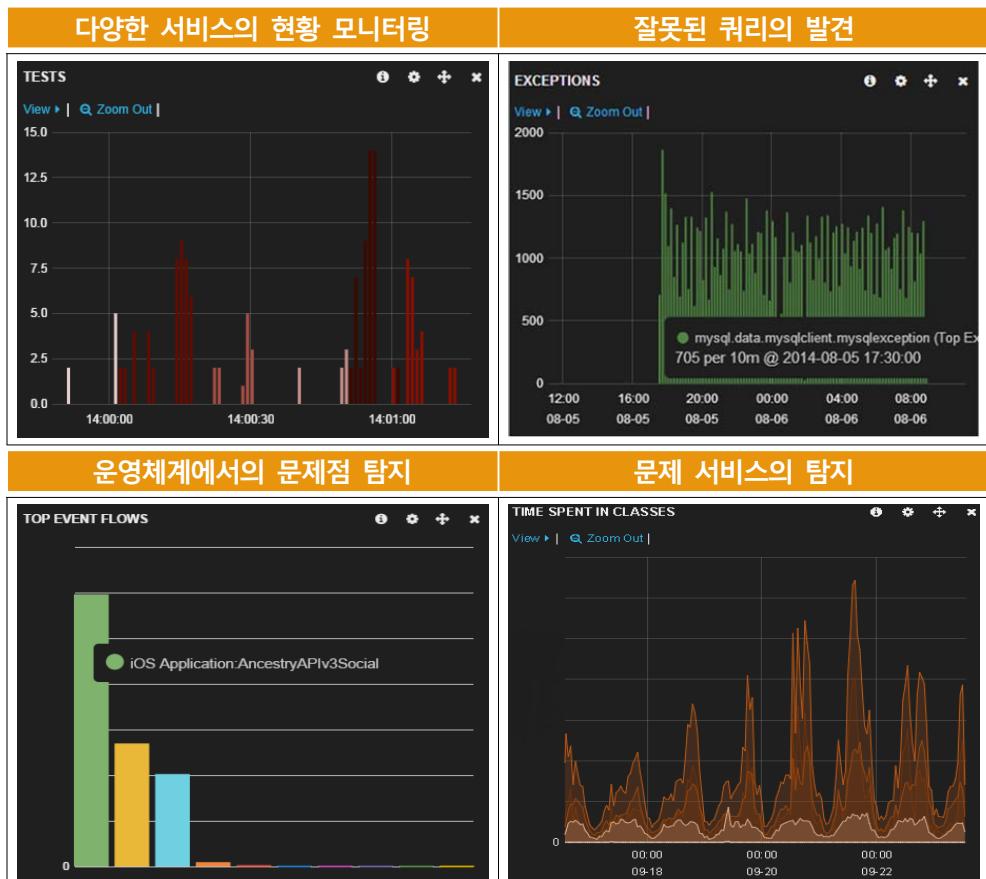
그래프로 표현된 가계도	DNA 매칭
<p>Family tree as a graph (DAG)</p>  <p>2020 nodes 572 marriage edges 2910 family edges</p> <p><a href="#">22</a> <a href="#">ancestry.com</a></p>	<p>DNA Matching</p> <ul style="list-style-type: none"> <li>Framing the Problem           <ul style="list-style-type: none"> <li>- 46 chromosomes, 23 pairs, 3 billion+ base pairs, AGTC</li> <li>- 99.9% of human DNA is shared</li> <li>- Academic programs work at small scale (GermLine)</li> <li>- <b>New samples are matched against all previous samples</b></li> </ul> </li> <li>Use Big Data technologies to create a scalable matching platform           <ul style="list-style-type: none"> <li>- Hadoop and MapReduce</li> <li>- HBase</li> </ul> </li> </ul> <p></p> <p>16 <a href="http://www.cs.columbia.edu/~gpari/gemini/">http://www.cs.columbia.edu/~gpari/gemini/</a> <a href="#">ancestry.com</a></p>



## ■ 고객서비스의 향상을 위해 서비스 사용현황과 고객 사용기록을 파악

- 현재 서비스의 사용 현황을 모니터링하여 특정 서비스 혹은 인프라의 문제점을 탐지 및 확인
- 어느 부분에 새로운 콘텐츠가 투입되고 만들어져야 하는지를 검색 행적을 통해 파악하여 콘텐츠 제공 방향을 결정하기 위한 자료로 활용
- 사용자의 불만족 시점이나 서비스 탈퇴 시점을 분석하여 고객 서비스 향상에 활용

[그림] Ancestry.com에서의 사용자 패턴분석



자료: Ancestry.com

- 사용자가 정당한 목적으로 웹사이트의 정보를 이용하는지 확인하기 위한 보안상의 목적으로도 활용

- 다양한 소스로부터 추출된 페타바이트급 데이터를 관리하기 위한 빅데이터 시스템 구축
  - 약 10PB의 데이터를 마이닝하고, 대량의 DNA 데이터를 다루기 위해서는 대규모의 데이터를 분산·병렬처리 하는 것이 효과적이라 판단
  - 서비스 데이터의 일부를 3개의 클러스터로 분리해 프로세스하고 있으며 하둡의 분산 처리기술을 통해 빠른 핸들링
  - 3개의 클러스터는 각각 DNA 매칭을 위한 데이터 마이닝, 머신러닝, 단순 데이터 구축을 위한 용도로 구성
  - 서비스의 중단 없이 지속적인 운영을 위한 고가용성이 매우 중요했으며 이에 MapR의 고가용성 JobTracker 활용
  - 이를 통해 다른 업무를 동일한 클러스터 상에서 처리하는 것을 가능하게 하였고, 시각적인 사용자 인터페이스와 클라이언트 구성 능력, 빠른 처리 등이 가능해짐

## 효과 및 향후 적용 확대 방안

- 향상된 디지털 이미지 처리 기술을 활용하여 정교한 사용자 데이터를 확보하고 모바일 특화 서비스를 개발
- 기업적 목적이 아닌 개인의 흥미를 위한 빅데이터 활용의 대표 사례로 개인의 성향, 취향, 미래 정보, 여행지 제안 등 비슷한 사례로의 무한한 확산이 기대
  - 개인이 알고 싶어 하는 맞춤화 되고 최적화된 검색 결과를 제시하기 위해서는 다양한 데이터 소스를 연계하고 축적하는 노력이 필수
- 이미 축적된 데이터뿐만 아니라 검색, 결제 등 개인의 활동을 통해 발생되는 사용 기록 데이터를 활용해 새로운 빅데이터 분석서비스를 제공

## 8. 오비츠, 사용자 특성을 파악하여 맞춤 검색 결과 제공



웹사이트로 유입되는 고객의 로그데이터를 파악하여 고객군을 분류하고 고객 특성별 다양한 호텔이 노출되도록 검색 결과를 조절하는 서비스 제공

### 추진 목적 및 배경

#### ■ 사이트 유입량 증가에 따라 고객 검색 데이터가 대량으로 축적

- 미국 온라인 여행 사이트인 오비츠는 항공권 및 기차표 예매, 호텔 예약, 여행 상품 정보 제공 등의 서비스를 제공하며 매일 최대 150만 건의 항공 검색과 100만 건의 호텔 검색 발생
- 이로 인해 일별 최소 500GB의 로그데이터가 발생되고 있으며 이러한 데이터의 저장과 처리를 위해 현재 데이터 인프라를 활용하는 것은 비싸고 어렵다고 판단

#### ■ 축적된 데이터를 효율적으로 저장하고 활용하고자 하는 인프라에 대한 요구 발생

- 오비츠가 보유한 대용량의 데이터 세트를 위한 스토리지 등 하드웨어 기반 마련
- 비용 효율적 운영 및 개발자와 분석가의 활용도를 높이기 위한 오픈 액세스 허용
- 실시간으로 데이터 쿼리를 처리하고, 어플리케이션 리포트를 신속하게 배치하기 위한 솔루션 필요

#### ■ 고객군 특성에 따라 구매력이 다르며 이를 검색 결과에 반영한 매출 증대 기회

- 직관적으로 맥OS 사용자가 일반 PC 사용자 보다 구매력이 높다고 추측하고 있었으나 실질적인 데이터로 이를 증명하고자 하는 호기심 발생
- 이를 통해 접속자 특성에 따라 다른 검색 결과를 보여주는 타깃 마케팅을 진행할 수 있는 기반을 마련하고자 함

## 추진 내용

### ■ 실질적인 마케팅 활용을 위해 인력 충원 및 분석팀 조직

- 빅데이터 경험이 있는 기업의 통계 전문가를 고용하여 새로운 분석팀을 만들었으며 대용량 데이터에서 유용한 정보를 찾아내는 데이터 마이닝을 우선 순위에 놓고 데이터간의 관계, 패턴, 규칙 등을 찾아내는데 주력

### ■ 데이터 마이닝을 통해 해당 사이트에 접속하는 PC(혹은 OS)의 종류에 따라 고객의 구매력의 차이가 있다는 결론 도출

- 맥 OS를 사용하는 사람의 구매력이 MS 윈도 OS를 사용하는 사람보다 30% 정도 높다는 직관이 실질적인 데이터를 통해 확인됨
- 맥 사용자는 PC 이용자보다 평균 20~30달러를 더 지출하는 경향이 있으며 4~5성급 호텔을 예약하는 비율도 40% 이상 높다고 분석

### ■ 윈도 사용자에게는 더 적은 비용의 모델을, 맥 사용자에게는 가격대가 상대적으로 높은 모델을 소개하는 등 맞춤화된 정보를 제공하기 시작(맥과 윈도우를 이용한 방문자에게 다른 호텔 옵션을 제시하는 것이며 같은 호텔을 다른 가격에 보여주는 것은 아님)

- 맥PC로 검색 시 일반PC 검색에서는 첫 페이지에 나오지 않았던 값비싼 부티크 호텔이 노출
- 숙박비가 더 비싼 일부 호텔은 두 경우의 검색 결과에 모두 노출되었으나 맥에서는 보다 상위 리스트에 노출되는 등 맥 검색의 첫 페이지에 나온 호텔들은 PC 검색의 첫 페이지에 나온 호텔들보다 약 11% 가격이 높은 호텔이 노출
- 아직까지는 이용자의 위치나, 호텔의 인기도와 홍보, 오비츠 사이트 등록일 등의 다른 요인들이 검색 결과에 더 크게 작용하나 향후 정교한 모델을 통해 PC(OS)의 차이도 검색 결과 영향력을 확대할 예정

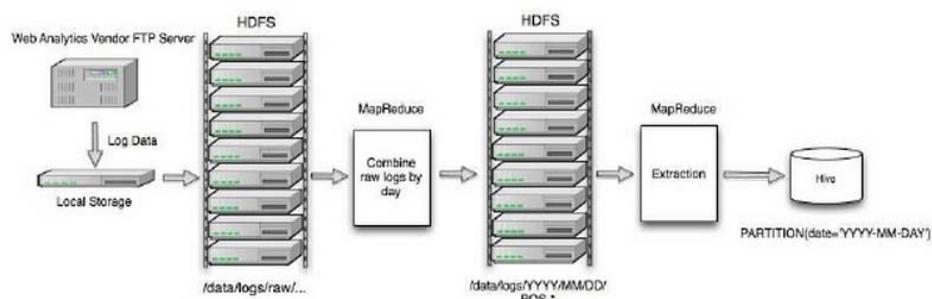
### ■ 하둡과 하이브를 활용하는 빅데이터 솔루션을 도입하기 시작

- 비용 효율화 및 신뢰성을 확보하기 위해 빅데이터 오픈소스 솔루션 도입을 추진
- 클러스터링 된 기기들 간 대용량 데이터를 다루는데 적합하며 확장이 용이한 HDFS(하둡 분산파일 시스템), 병렬적으로 연결된 대량의 데이터를 분산 처리하는데 효과적인 맵리듀스, 오픈소스 데이터 웨어하우징 솔루션 하이브를 적용

### ■ 검색 프로세스 개선을 통해 빠른 호텔 검색결과 노출 등 개선 작업 착수

- 맵리듀스 프로세싱을 위해 웹트렌드의 로그데이터로부터 데이터를 발췌하여 기존 프로세스 소요시간이 약 100분 걸렸던 것 대비, 맵리듀스 프로세스 소요 시간은 약 25분으로 감소
- 하이브를 통해 이전에는 불가능했던 작업을 쉽게 처리함. 예를 들어 검색결과에서 각각의 예약된 호텔의 현황을 찾는 것이 가능해졌으며, 위치나 일수 기준으로 예약 현황을 종합할 수 있음

[그림] 웹 애널리틱스 데이터의 처리 프로세스



자료: Orbitz Worldwide

## 효과 및 향후 적용 확대 방안

- 고객의 온라인 활동을 추적하여 특성을 파악하고 이를 통해 고객 기호나 지출 습관을 예측하여 최적화된 제품/상품을 제시하는 것이 가능해짐
  - 오비츠의 사례를 시작으로 고객군 특성별 구매력 차이가 발생하는 지에 대한 실험이 다양하게 제시되어 마케팅에 적극적으로 활용될 예정
  - 특히 온라인 기업들은 호텔/여행 분야 외 다른 분야에서의 맥 이용자들과 나머지 이용자들의 구매력 차이에 대한 분석에 관심을 갖고 있음
- 데이터를 통해 고객의 미래 쇼핑 습관을 예측하고 잠재 고객을 파악하는 등의 ‘예측 분석(predictive analytics)’ 확대
  - 고객 데이터를 통해 성향을 분석하여 타깃 마케팅을 함으로써 매출 증가가 가능하다는 좋은 사례로 다양한 온라인 비즈니스에 확대 예측
  - 온라인 기업들은 그들의 제품과 서비스에 대한 가장 높은 “생애 가치(lifetime value)”를 갖는 대상을 위한 맞춤화 서비스 제공

## 9. NC소프트, 게임 내 사기 탐지 시스템 구현



게임 사용자들이 생산하는 방대한 양의 로그기록을 이용하여 회귀 분석, 자기 유사도 알고리즘 및 기계 학습을 통해 게임 버그와 비정상적 사용자 탐지

### 추진 목적 및 배경

#### ■ 온라인 부정 거래, 사기 행각의 심각성 확대

- 다른 사람의 자산을 불법적으로 탈취하는 행위, 즉 신용카드 도용, 은행계좌 도용, 보험 사기, 탈세를 종합하여 'Fraud'라고 함. 이런 사기 행각들은 전체 온라인 거래의 9%를 차지하며 거래 비율은 매년 약 2배씩 성장하는 추세 (Online Fraud Report, CyberSource, 2012)
- 방대한 거래 데이터를 다루는 기업들은 사기 탐지(Fraud detection)에 다양한 데이터 분석도구를 이용하고 있으며 정확한 알고리즘을 생성하기 위해 금융 회사, 통신회사, 결제 대행회사들은 상당한 투자를 집행하고 있음

#### ■ 게임 사기(Game Fraud) 탐지 - 게임 내 아이템의 불법적 거래 및 사기 급증

- 엔씨소프트의 주력 게임인 리니지, 리니지2, 아이온, 블레이드 앤 소울 등의 MMORPG<sup>4)</sup>에서는 실생활과 유사한 생산 및 소비 행위가 발생하고 있으며 게임 내에서 생산한 가상 재화를 현금과 맞교환하는 블랙 마켓이 크게 활성화되어 있음. (게임아이템 현금 거래와 정보보호, 한국정보보호진흥원, 2006)
- 소위 '오토'라고 불리는 게임 자동 사냥 프로그램을 통해 경험치 및 재화를 손쉽게 취득하고 이를 블랙마켓을 통해 현금화하는 전문적인 사업자들이 증가

4) MMORPG : 대규모 디중 사용자 온라인 를 플레이 게임(Massive Multiplayer Online Role Playing Game)의 줄임말. 게임 속 등장인물의 역할을 수행하는 형식의 게임인 RPG(롤 플레이 게임)의 일종으로, 온라인으로 연결된 다수의 사용자가 같은 공간에서 동시에 즐길 수 있는 게임을 말함(네이버 백과사전)

및 불법 기업화되고 있어 이에 대한 발빠른 대응이 중요해지면서 게임 내 정상 사용자와 불법 사용자를 구별하고 탐지해 내는 기법이 필요해짐

- 이 외에도 다른 사람의 계정을 도용하여 보유한 게임 아이템 및 기타 자산을 몰래 처분하는 행위나 게임의 버그를 악용하여 몰래 아이템의 무한 복제, 보스몹 무한 사냥 등을 수행하는 등 불법적인 행동이 나타나 선량한 사용자의 직접적인 피해는 물론 정상적인 게임을 방해하는 등의 문제가 심각해지면서 이에 대한 단속이 필요

## 추진 내용

### ■ 보다 높은 탐지율을 위해 빅데이터 분석 기법 도입

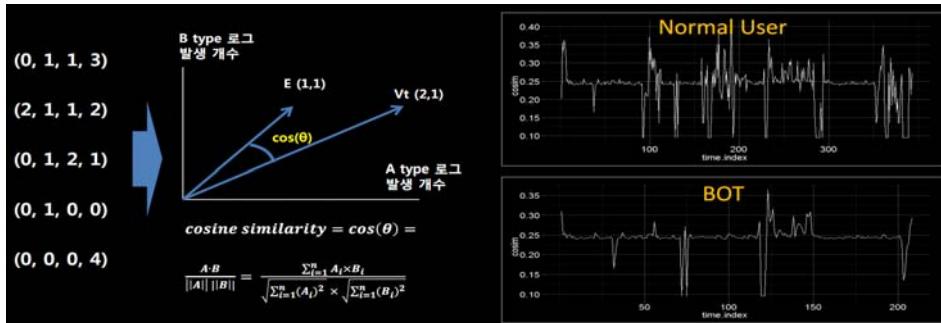
- 수많은 게임 사용자의 플레이를 통해 발생되는 대량의 로그 데이터를 수집 / 적재하고 이를 가공하기 위해선 빅데이터 처리 기술이 요구되며 가공된 데이터를 통해 탐지 패턴을 찾아내기 위해선 통계 분석 및 기계학습을 이용한 탐지 모델 필요
- 로그 데이터 적재 및 관리를 위해 하둡 클러스터를 구축하고 데이터 가공은 Pig와 Cascading, 분석 및 모델링은 R을 이용

### ■ 오토 캐릭터의 탐지 - 자기유사도 알고리즘 + 로지스틱 회귀분석

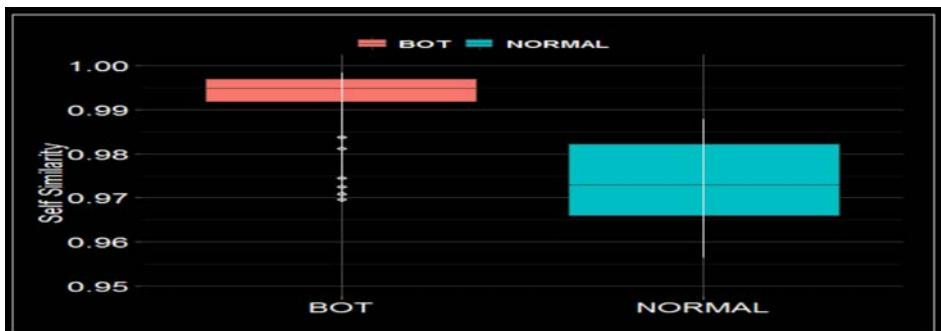
- 오토 캐릭터는 미리 설정된 행위를 반복하는 경향이 강하기 때문에 각 캐릭터별 자기 반복적인 경향을 정량화하고 이렇게 정량화된 수치가 높은 캐릭터들을 탐지하는 것이 핵심 요소

\* 정량화된 수치는 '자기 유사도 알고리즘(Self similarity Algorithm)'로 정의하며 오토 캐릭터 탐지에 가장 핵심이 되는 요소

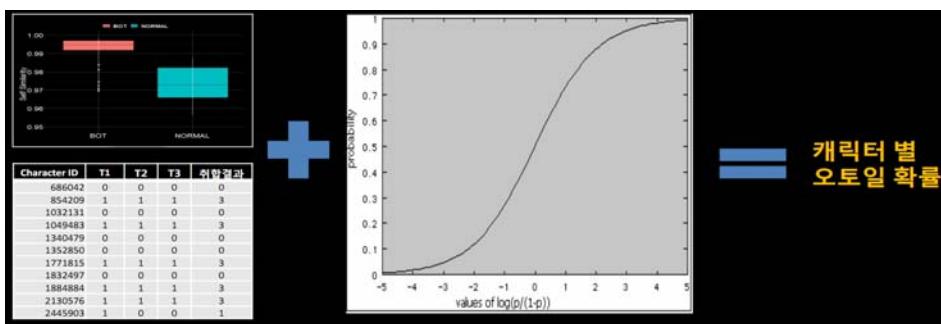
[그림] 캐릭터 별 발생 로그를 벡터를 변환 → 각 벡터들의 코사인 유사도(Cosine Similarity)를 계산



[그림] 캐릭터 별 코사인 유사도 표준 편차 계산 후 자기 유사도 값으로 변환



[그림] 정답 집합을 이용하여 자기 유사도 값을 BOT 확률로 전환



자료: NC소프트, 2014

## ■ 뱅커(Banker) 캐릭터의 탐지 - 네트워크 분석(Network Analysis)

- 기존 탐지 시스템으로는 은밀한 위치 오토 캐릭터가 수집한 경험치와 재화를 통합관리하는 뱅커의 검출 및 탐지 작업을 수행하는데 한계가 나타남
- 사용자들의 방대한 행동을 담은 로그 데이터 분석을 통해 캐릭터들 간 거래 네트워크를 구성한 후, 그래프 클러스터링을 수행하여 오토 캐릭터가 많이 활동하는 클러스터(작업장)을 구분 한 후 뱅커 캐릭터를 탐지하는 로직 개발

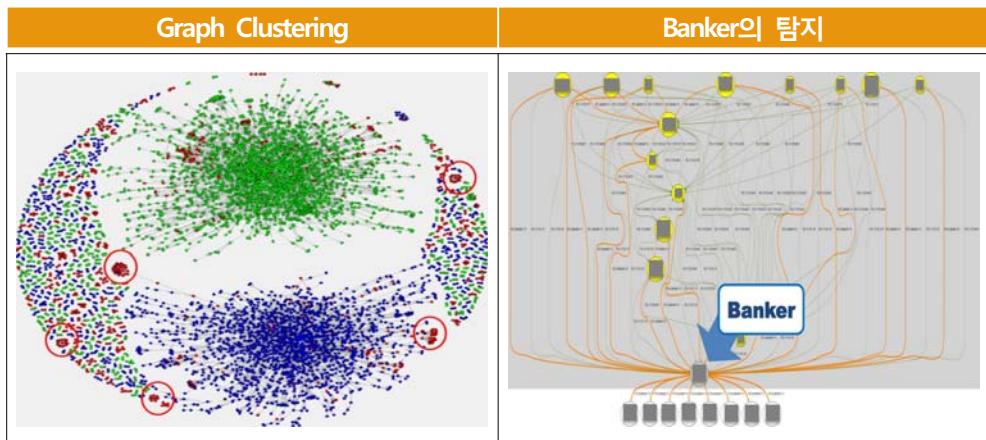
[그림] 뱅커 탐지를 위한 사용자들 간의 관계 네트워크 시각화



자료: NC소프트, 2014

- 그래프 클러스터링(Graph Clustering): 긴밀한 네트워크 형성 집단을 분류하고 작업장 여부 판별 기준을 적용

[그림] 그래프 클러스터링 및 Banker의 탐지 시각화



자료: NC소프트, 2014

- 뱅커 캐릭터를 찾아 자산 앱류 등 작업장에 실질적인 경제적 타격을 가함으로써, 불법 사용자를 감소시키고 게임의 정상적 운영을 꾀하는 효과를 보임

## 효과 및 향후 적용 확대 방안

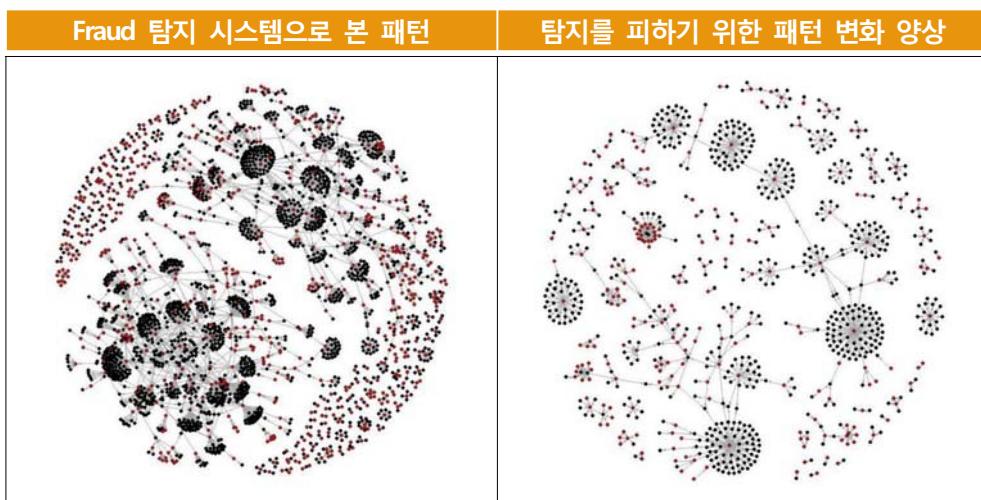
### ▣ 오탐률 감소에 대한 숙제 ‘여전’

- 일반 정상 사용자를 오토로 잘못 판단하거나 새롭게 게임을 시작하는 친구에게 좋은 아이템을 선물했는데 이를 앱류당하는 사례 등 부정 사용자에 대한 알고리즘 불완전성은 여전히 숙제로 남아있음
- 해외에서 다양한 표적 알고리즘 사례가 소개되고 있으나 오탐률이나 잘못된 결과에 대한 내용은 공유되고 있지 않아 더 나은 알고리즘 개발 속도는 다소 더디게 진행된다는 점이 문제

■ 사기 행위자들은 탐지 패턴을 피하기 위해 끊임없이 ‘변화’

- 오토 및 사기 관련 불법 캐릭터들은 지속적인 정보 공유와 대응 방법 연구를 통해 탐지 패턴을 피하기 위해 변화하고 있으며 게임사와 대결 구도에 있음
- 특히 탐지 효과가 클수록 더 적극적으로 패턴을 변화시키며 발전하는 추세

[그림] 탐지 시스템에서의 패턴에 대응하는 Fraud 패턴 변화 양상



자료: NC소프트, 2014

■ 게임데이터는 활용성 및 가능성성이 매우 높은 데이터

- 사기탐지를 비롯한 게임데이터는 향후 활용 가능성이 유용한 데이터지만 자료 공유가 거의 없어 분석 알고리즘 및 관련 데이터 분석에 한계가 있음. 따라서 게임사끼리의 자료 공유가 활발하게 나타나야 하며, 게임 이외의 데이터 결합, 분석을 통해 더 큰 사회적 가치 창출이 가능함



### ■ 오픈 소스 기반 데이터 플랫폼 적용 확대

- 보다 실질적인 가치 창출을 위해서는 지속적으로 활용성 및 효율성을 극대화할 필요가 있음. 이를 위해선 대규모 데이터 처리를 위한 하둡과 같은 빅 데이터 플랫폼 구축이 보다 확대되어야 하며, 다른 개발툴과의 연결 및 시스템 연동에 유리한 R의 활용이 점차 중요해질 것으로 전망
- 특히 빠른 탐지 및 대응을 위해 실시간 데이터 처리 인프라 구축 및 적용을 검토하고 있음

## 10. 멜론, 이용자 관심도에 따른 콘텐츠 추천



사용자들이 축적한 데이터를 통해 아티스트별 인기도를 측정하고, 이용자 관심사에 맞는 흥미유발 콘텐츠를 추천하는 음원 서비스 제공

### 추진 목적 및 배경

- 로엔 엔터테인먼트의 음악 서비스인 멜론은 국내 2,400만 이용자를 보유하며 320만 음원을 보유하는 대규모 서비스로 그동안 축적된 데이터를 활용하고자 하는 과제 당면
- 기존 로엔 엔터테인먼트의 관계형 데이터베이스로는 현재 멜론의 방대한 데이터 관리에 한계에 도달
  - 일평균 7천만 건 이상의 스트리밍이 발생하며, 월 평균 1,200만 이상의 방문자 유입, 연간 10억 건(하루 320만 건)이상의 콘텐츠 이용이 발생
  - 관계형 데이터베이스를 이용한 소식 서비스의 부하 발생을 계기로 멜론의 빅데이터를 수용할 수 있는 적절한 솔루션 모색을 시작

### 추진 내용

- 멜론의 적절한 데이터 관리와 소비자 이용 경험 증대를 위한 빅데이터 필요성 증가
  - 기존 배치 애플리케이션의 한계 및 다양한 분석 알고리즘이 부재한 상황
  - 방대한 양의 데이터를 적재 및 보관하고, 분석하고, 재사용하기 위한 대용량의 하드웨어와 소프트웨어가 필요한 시점
  - 이용자 관점에서 빠른 검색 및 조회를 할 수 있는 기반 마련 필요

### ■ 멜론 이용자들이 축적한 데이터를 통해 팬 소비지수 개발

- 이용자의 음원 소비량, 영상 재생횟수, 콘텐츠 조회수, 콘텐츠 좋아요 수, 콘텐츠 공유 수, 댓글 등 이야기 수 등 다양한 활동 데이터가 멜론에 축적되어 있었음
- 이러한 데이터 중 대표적인 31가지를 선택하여 분석한 뒤 이용자들의 활동을 점수로 환산
- 이용자 활동 점수를 바탕으로 아티스트별 팬 선호도를 측정
- 멜론에 등록된 콘텐츠에 대한 이용자 반응을 일별, 주간, 월간 기준으로 파악 가능

### ■ 특정 아티스트를 선호하는 팬을 대상으로 타깃 마케팅 진행

- 팬소비지수를 통해 가장 팬 선호도가 높은 아티스트를 파악하거나 개별 아티스트가 보유한 팬들과 그들의 관심 정도 등의 관계 데이터를 추출 가능
- 예를 들어 A라는 아티스트의 팬, 매니아, 잠재팬의 분포와 그들의 성별, 연령별 분포, 아티스트에 대한 조회수와 공유수, 좋아요 수를 파악 가능
- 이를 기반으로 아티스트의 팬이나 잠재 팬 대상으로 소식을 발송하여 관심도를 높이고, 팬을 증가시키고 장기적으로 멜론 음원 서비스의 이용 확대가 일어날 수 있는 선순환 구조를 구축

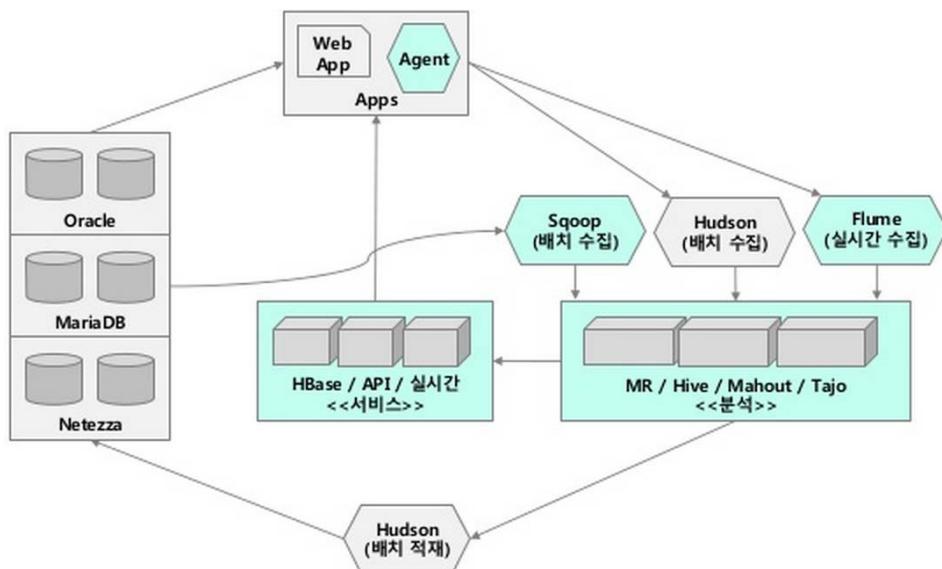
### ■ 이용자의 관심을 유도하기 위한 친밀도 분석 결과 제시

- 이용자가 특정 아티스트와 관련하여 발생시킨 데이터양을 기반으로 하여 이용자의 아티스트에 대한 친밀도를 온도(°C)로 표현하여 제시해 주며, 수많은 팬 중에 나의 순위를 숫자로 제시하여 흥미 유발
- 이용자에게 흥미와 즐거움을 유발함으로서 아티스트에 대한 충성도를 높이거나 새로운 아티스트에 대한 관심을 높이도록 자연스럽게 유도하며 이는 멜론 음원 소비에도 긍정적으로 작용

## ■ 테라바이트 단위의 방대한 데이터를 저비용으로 관리하기 위한 솔루션 선택

- 비용이 높은 상용 솔루션 보다 저렴하면서 안정적인 대용량 데이터 서비스를 제공하는 오픈소스 솔루션을 활용하기로 결정
- 검증을 통해 최종적으로 하둡, HBase, Mahout을 선택하였고 기술 내재화를 위해 파트너사를 선정하여 내부 인력 문제를 해결

[그림] 멜론 데이터의 수집-분석-서비스 아키텍처 구성



자료: 로엔 엔터테2014

- 일평균 2TB 이상의 데이터가 멜론 및 이용자에 의해 생성되고 있으나 최소한의 정보만 적재하는 것을 원칙으로 하고 있으며 현재 약 300TB의 정보를 마리아 DB로 관리 중
- 일부 서비스의 경우 발생시점에 따른 관리 정책을 사용하여 폭증하는 데이터를 적절히 조정하고 있으며 비용편의 검토를 통해 데이터 적재를 최소화

## 효과 및 향후 적용 확대 방안

### ■ 로엔 엔터테인먼트의 빅데이터 도입의 핵심은 기술 내재화

- 하둡 등 오픈소스 솔루션의 국내 인력 품귀로 내부 인력 부채 상황. 그러나 장기적 운영을 위해서는 지속적인 학습이 가능한 인력이 필요
- 이에 기술 내재화를 위한 파트너사를 선정하였으며 기술 중심의 빅데이터 도입이 아닌 비즈니스 중심을 유지하기 위해 장기적인 시간을 두고 준비
- 사내 서비스 기획팀과 기술개발팀 간 1차 커뮤니케이션을 통해 목표를 명확히 하였으며, 기술개발팀과 파트너사(그루터)를 통해 하둡 플랫폼 기술 지원 및 운영 지원을 받아 진행

### ■ 향후 목표는 이용자의 활용 패턴을 이해하여 특정 아티스트 추천이나 맞춤형 서비스를 먼저 제안하는 프로그램을 개발하는 것

- 빅데이터 서비스가 확산될수록 개인 맞춤형 제품/서비스 추천 기능이 점차 확산되며 정교해 질 것이며 멜론 서비스도 이에 적극 대응 예정
- 고객 이력 기반 추천, 콘텐트 기반 추천, 메타 기반 추천을 활용할 계획
- 장기적으로는 스마트카, 웨어러블 디바이스와 연동한 지능형 서비스 개발 예정



# III

## 의료





## 11. UNC 헬스케어, 환자의 재입원 비용 절감



텍스트 분석 기술을 통해 저소득층 대상의 의료 보장제도 하에 있는 저소득층 환자들의 재입원 비율을 줄여 의료비용 절감

### 추진 목적 및 배경

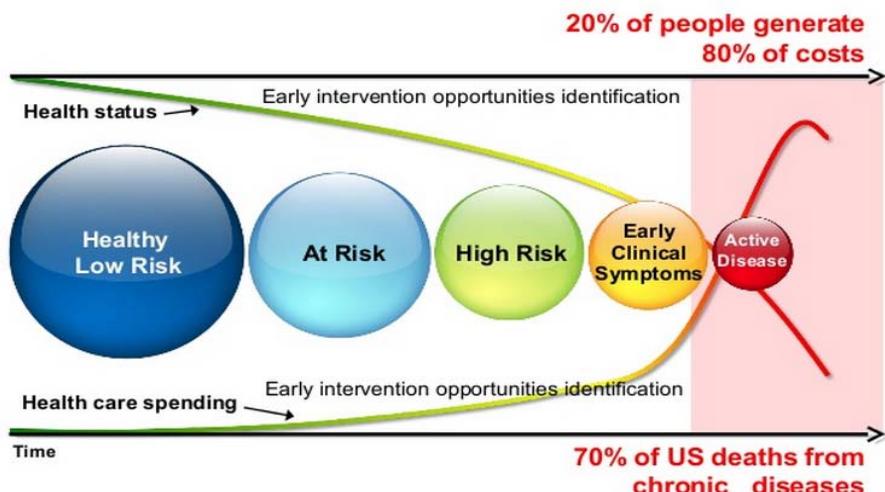
- 저소득층 대상의 의료보장제도의 등록자 수가 증가
  - 경기침체 이후 2012년 기준 약 700만 명이 저소득층 의료 보장제도 등록

#### ※ UNC 헬스케어

UNC 헬스케어(University of North Carolina Health Care)는 비영리 통합 의료 기관으로 노스 캐롤라이나 북부에 설립되어 있고, 채플 힐에 본사를 두고 있다. UNC 헬스케어는 UNC-채플힐 의학대학과 전국적으로 저명한 연구기관과 연계되어 있고 매년 37,000명 이상의 환자를 수용하는 최첨단 시설로 고품질의 의료 서비스를 제공한다.

- 저소득층 의료보장제도에 대한 국가 비용문제 발생
    - 저소득층 의료보장제도의 자금은 미국 연방정부와 각 주에서 공급
    - 저소득층 의료보장제도의 높은 비용과 비효율적 운영으로 인한 경제적 피해로 비용을 줄이고, 효율적인 운영이 필요하다고 판단
- ※ 20%의 사람이 80%의 질병비용을 일으키는 파레토법칙이 적용

[그림] 질병 위험 상태와 병원 지출 비용의 관계



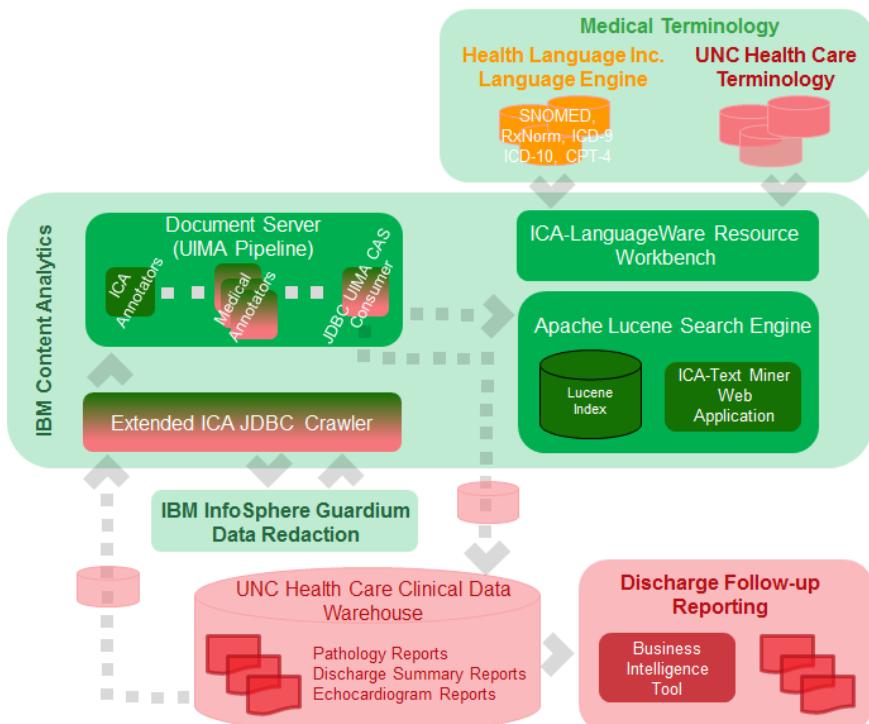
자료: IBM, 2014

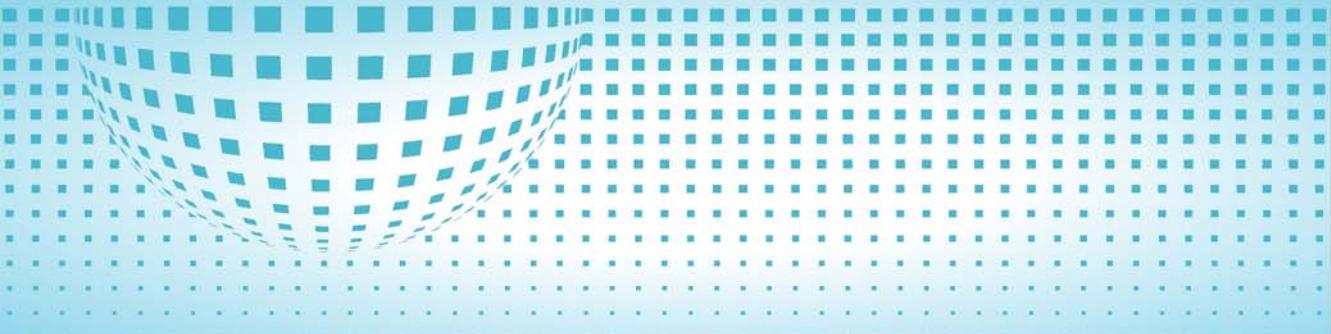
- 의료데이터의 대부분을 차지하는 비정형화된 데이터의 분석 필요성 증가
  - 의료기관은 수많은 데이터가 존재하며 신뢰성이 확보된 데이터 및 정보에 대한 통찰력을 확보할 필요가 생김
  - 의료 데이터의 80%이상이 비정형화된 데이터로 기존 인프라를 통한 분석이 어려웠으며 이를 해결하기 위한 서비스 도입이 필수적인 상황
- 타 의료기관에서의 암진단 자료와 같은 외부자료의 연계가 불가능한 경직된 정보 시스템 구조로 원활한 진료가 어려운 환경

## 추진 내용

- 유방 촬영술(mammography screenings)과 자궁경부세전 검사(Pap Smear)에 있어 콘텐츠 분석(Content Analytics)과 자연어 처리(Natural Language Processing)를 활용한 비정형 의료 데이터 분석 수행
  - 환자의 영상과 텍스트 데이터에서 비정상(Abnormal) 정보를 추출해 내는 일은 의료진의 많은 시간과 노력을 필요로 함. 기계 판독과 자동처리 알고리즘을 통해 비정상 부문을 자동 추출하여 의료진의 시간과 노력을 절감

[그림] UNC 의료정보 분석 플랫폼





## ■ 데이터 분석을 통한 통찰력 확보를 위해 IBM의 텍스트 분석 서비스를 도입하였으며 환자 재입원 비용을 줄이는데 활용

- 정형 및 비정형 데이터를 모두 확인 가능해졌으며 특히 비정형화된 데이터를 통해 환자 입원 예방 조치에 활용
- 환자들의 입원 원인에 대한 파악이 가능해짐에 따라 입원을 방지하기 위한 예방 조치도 확립 가능해짐

## ■ 가공되지 않은 정보를 사용 가능한 정보로 변환하는 것을 통해 의료산업에 대한 다양한 의사결정 및 통찰력을 마련함

- 재입원을 줄이기 위해서는 입원에 위험이 보이는 환자에게 적시에 의료 서비스를 제공해야 하고, 입원을 했던 환자가 병원 퇴원 후 문제가 있을 경우 빠른 후속 조치가 필요
- 기존에는 이러한 환자 정보가 비정형 문서로 저장되어 있기 때문에 확인이 어려웠으나 빅데이터 분석 솔루션 도입 후 이러한 비정형 데이터를 처리/분석하여 환자들의 문제점을 파악 가능
- 또한 문제점을 발견하는 시간도 줄어들어 환자의 재발병을 미리 인지하여 빠른 후속조치를 제공하는 것이 가능해짐
- 이러한 프로세스를 통해 의료보장제도 혜택을 받는 저소득층 환자의 재입원 비율 및 그에 따른 비용을 줄일 수 있음

## ■ 환자의 쉬운 열람이 가능하도록 데이터를 변환

- 건강관리에 대해 환자가 직접 참여하게 하는 것은 건강회복을 위해 중요하나 의료 데이터는 이해하기가 어렵고 필터링 되지 않았다는 문제점 보유
- 의료 데이터를 단순한 형식으로 변환시켜 환자들이 자신의 건강 정보를 이해하기 쉽도록 하였으며 이를 통해 건강관리에 직접 참여할 수 있게 함

## 효과 및 향후 적용 확대 방안

- 빅데이터를 활용해 유방암과 자궁경부암 부문에서 암진단 건수를 10% 이상 증가 시켰으며, 결장암(colon cancer)과 같은 타 암의 진단에 확대 적용

[표] 빅데이터 적용성과(IBM, 2011)

구분	전반적 정확도	정밀도	민감도(리콜)	특이도	양성예측도
진단	78%	90%	80%	68%	90%
사후관리	79%	95%	74%	91%	95%

- 빅데이터를 적극 활용해 선진화된 의료 서비스를 갖추고, 맞춤형 건강관리 프로그램과 같은 다양한 환자 관리 프로그램을 마련하는 등 다양한 빅데이터 서비스 확산
- 의료진과 환자 간의 소통 및 의료기관간 데이터 교환, 안전한 클라우드 컴퓨팅 인프라에 기반하여 작은 의료기관들이 사용할 수 있는 의료정보 운영환경 제공
- 방위 의료 및 임상 데이터 확보를 통한 임상 연구 역량 강화는 물론, 전체 의료 데이터의 통합 분석을 통한 의료 서비스 품질 향상에 도움을 주며 장기적으로 업무 효율성 강화 및 데이터 투명성 강화에도 기여할 것으로 기대
  - 민감한 환자의료 정보를 활용하거나 공유하기 위해서는 익명화, 암호화 등 보안문제해결이 필수적

## 12. 서울아산병원, 의료연구 편의성 확대



대용량비정형 의료 데이터를 효과적으로 암호화하고, 법규준수를  
이행하는 의료 연구 목적의 연구정보검색시스템 개발

### 추진 목적 및 배경

#### ■ 개인정보 보호법 강화로 다양한 분야의 의료 정보 보호에 관심 증가

- 진료나 경영목적으로 활용하는 개인정보 외에 의료진이 연구목적으로 수집 및 분석하는 정보에 대한 보호에 관심을 갖게 됨
- 개인정보보호 관련 다양한 규제 등장 속에서 의료 정보 시스템뿐 아니라 임상 연구에 쓰는 데이터에 포함된 개인정보의 유출 및 오남용 방지 대책 수립이 필요
- 생명윤리 및 안전에 관한 법률과 개인정보 보호법 모두를 충족시키기 위한 방안 마련

#### ■ 실무 차원의 규제 대응 가이드라인 마련과 시스템 설계가 어려운 실정

- 개인정보 보호법의 경우 모든 업계에 적용되는 보편성을 띠고 있으며, 생명 윤리 및 안전에 관한 법률의 경우 의료계를 위한 규제
- 반면 이러한 의료 데이터의 암호화나 익명화 등 기술적 보호 조치에 대한 구체적 가이드라인이 없어 실무자에게 혼란을 초래하므로 실무 차원의 규제 대응 지침을 마련하고 시스템을 설계할 필요가 있음

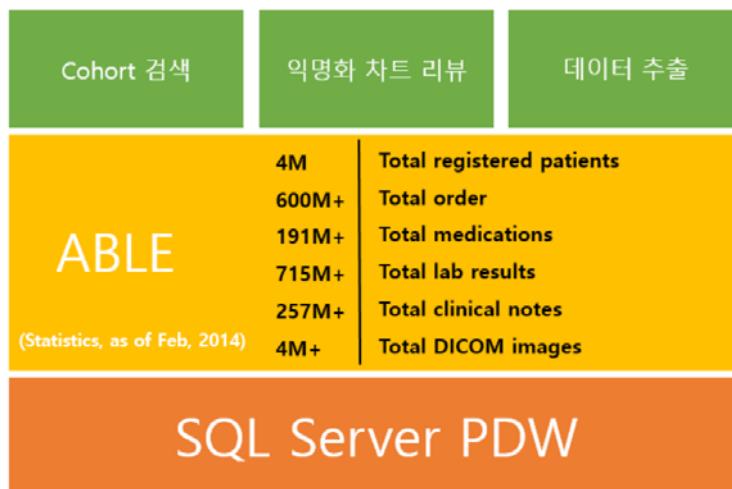
#### ■ 신속하고 정확한 연구 데이터 활용에 대한 요구 발생

- 기존에는 연구를 위해 원하는 데이터를 받아보려면 평균 일주일에서 한두 달이 소요되었으나 개선을 통해 즉각적으로 검색결과를 받아볼 수 있으며 쉬운 활용이 가능하도록 변경

## 추진 내용

- 익명화 처리기능과 대규모 정보 핸들링에 최적화된 데이터 플랫폼인 ABLE(Asan BiomedicaL research Environment) 시스템 구축 결정
- 의료 빅데이터를 활용하기 위한 핵심 기술인 비정형 데이터에 대한 익명화 성공 여부를 철저하게 검증
  - 일반적인 마스킹 솔루션으로는 비정형 데이터와 영상정보를 대상으로 한 익명 처리가 어렵다는 한계 보유
  - 이에 따라 ABLE의 익명화 모듈에 대한 개념 검증을 미리 실시하고 테스트를 통해 대규모 데이터 환경에서 익명화가 필요한 정보를 빠르고 정확하게 처리 할 수 있다는 결론 도출
- 이러한 내부 검증 결과를 바탕으로 본격적인 빅데이터 시스템 구축 작업 착수
  - 우선 데이터 처리를 위한 하드웨어 인프라를 구성하고, 빠른 데이터 처리를 위한 최적화 구성 등 성능 확보 작업을 실시하였고 비정형 데이터를 실시간 으로 익명화 처리가 가능하도록 고성능의 인프라를 구성
  - 빅데이터 어플라이언스가 핸들링 하는 데이터 규모는 6억건 이상의 오더 정보, 7억 2천만 건 이상의 검사 정보 등을 포함한 4백만 명의 환자정보 저장(2014년 기준)

[그림] 아산병원의 빅데이터 인프라가 수용하는 데이터 규모



자료: Microsoft, 2014

- ABLE 솔루션이 제공하는 주요 서비스는 크게 코호트(Cohort) 검색<sup>5)</sup>, 익명화 차트 리뷰<sup>6)</sup>, 자료 추출<sup>7)</sup>로 구성되며 핵심 기술인 익명화를 위한 마스킹 처리 구현
  - 보호 대상 개인정보를 정의한 후 구조화된 데이터의 경우는 삭제 조치를 하고, 기록지나 영상에 적혀 있는 구조화되지 않은 정보는 마스킹 기법으로 익명화
  - 텍스트의 경우 정규식 표현률을 적용하였고, 의료 영상정보는 DICOM(Digital Imaging and Communications in Medicine) 표준 활용
  - 추가적으로 ABLE 사용자 화면 구성과 데이터 전달을 위한 시스템을 개발

5) 코호트(Cogort) 검색: 의료진이 연구 가능성 검토를 위해 연구 대상 집단에 대한 조회를 위해 사용하는 서비스

6) 익명화 차트리뷰: 연구 대상군에 대한 세부적인 항목을 익명화된 상태로 조회하기 위한 서비스

7) 자료 추출: 의료진이 분석에 활용할 수 있게 필요한 데이터들을 추출할 수 있는 서비스

## 효과 및 향후 적용 확대 방안

- 연구 정보 획득 시간을 획기적으로 줄여 의료 연구 프로세스를 개선
  - 연구용 자료 신청 후 데이터를 받기까지 기다려야 했던 시간 없이 개인 컴퓨터상에서 바로 결과가 출력되고, 필요한 정보는 다운 받아 분석할 수 있는 등 실시간성 강화
- 의료 개인정보보호에 대한 의학계의 글로벌 표준 적용
  - 서울아산병원의 ALBE은 개인정보 보호법과 생명윤리 및 안전에 관한 법률 준수를 위해 글로벌 스탠더드를 적용한 국내 최초 사례로 평가

## 13. 맞춤형 유의질병 및 병원정보 제공



진료정보 빅데이터 분석을 통해 발생 질환별 예상 유의 질병 정보 및 맞춤형 병원 정보 제공

### 추진 목적 및 배경

#### ■ 사업 추진의 배경

- 정부기관 및 지방자치단체가 보유한 데이터를 누구나 손쉽게 활용하고 이를 통해 새로운 가치를 창조하고자 하는 '정부 3.0'의 방침에 따라 접근이 어려웠던 대규모 보건의료 공공데이터가 개방



#### ■ 사업 추진의 필요성

- 진료/질병/처방 등의 데이터가 지속적으로 생성되고 있는 보건의료 분야의 방대한 데이터를 활용하여 국민건강 증진을 위한 서비스를 제공하고자 함

- 또한 Open API를 통해 빅데이터 분석 결과를 외부에서 참고 가능하도록 하여, 정부 3.0 및 창조경제 부합하는 새로운 고부가가치 서비스를 창출하고자 함

## 추진 내용

| 참여기관 | 메디벤처스(주), 건강보험심사평가원, (주)라인웍스

| 주요 활용데이터 | 건강보험심사평가원의 건강보험청구 데이터베이스 등 60여종의 데이터를 기반으로, 다양한 공공 개방 데이터를 수집 활용함

※ 개인정보를 비식별화하여 분석

구분	데이터	데이터 양	제공 기관
건강보험청구 데이터	건강보험청구 데이터베이스 등 60여종 (요양급여비용청구명세서 / 요양기관정보 / 의약품처방정보)	50 TB	건강보험심사평가원

## | 분석 내용 및 기법 |

### 데이터 처리 및 분석 기법

#### ■ 유사그룹 기반 데이터 선별 (Collaborative Filtering)

- 유사그룹 기반 데이터 선별은 가상의 사용자 그룹을 생성하고, 그 안에서 가장 빈도수가 높은 데이터를 검색하는 기법
- 건강보험청구 데이터에서 주상병-부상병의 연결성을 정렬하여 데이터 분석 기계학습(Machine Learning)을 위한 학습데이터를 생성

## ■ 질병기준 인구통계학적 데이터 선별 (Demographic Curating)

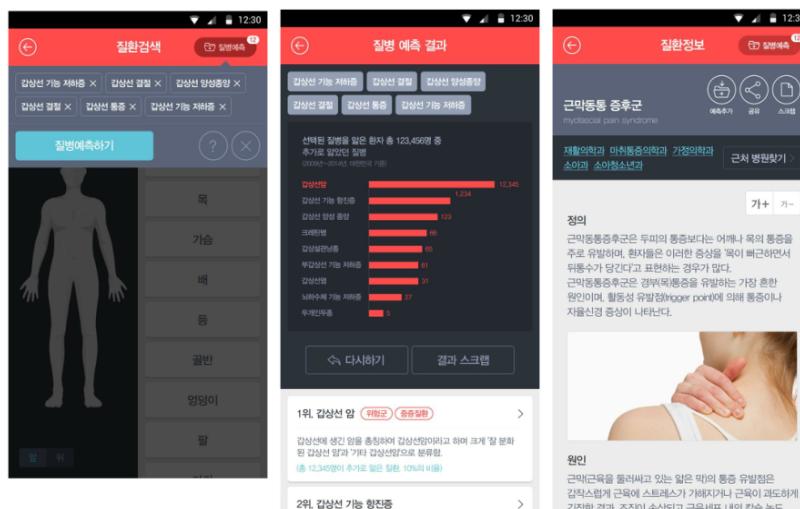
- 인구통계란 사용자의 부가 속성, 즉 나이, 성별, 거주지 등의 일반적인 정보를 의미하며, 이러한 인구통계학적 데이터와 질병은 연관도가 매우 높아 유행성 질병군의 예방 및 치료를 위한 데이터 추출에 활용
- 서비스를 통해 입력 받은 인구통계학적 정보를 데이터 선별의 기준 조건으로 카이스퀘어 분석(Chi-Square Test)을 통해 속성과 질병과의 연관성을 분석

## 주요 분석 결과 및 활용방안

### | 주요 분석 결과 |

#### ■ 국민질환 통계자료를 기반으로 개인 발생 질환별 유의질병 정보 제공

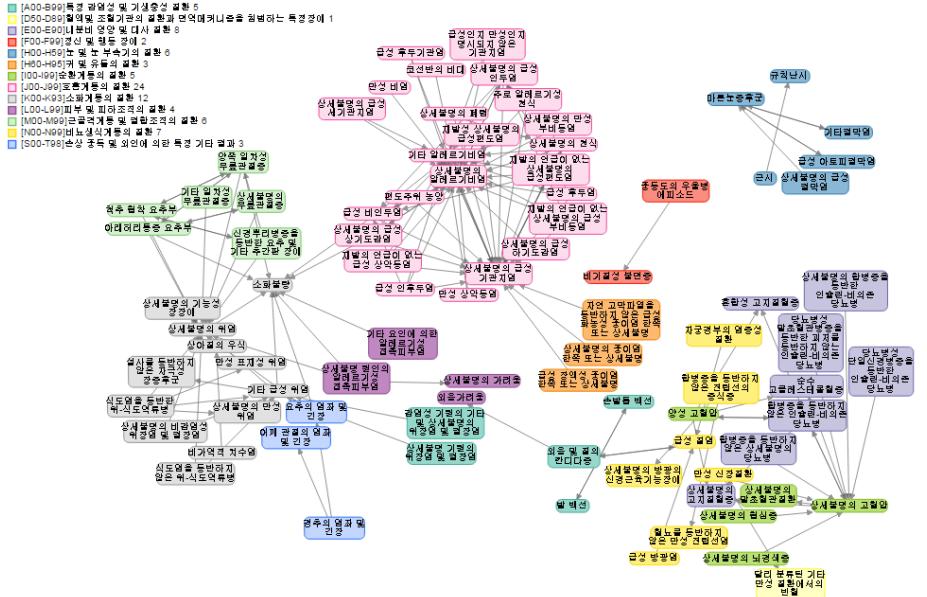
- 예) '치아 우식증(충치)' 환자의 경우 '위염' 발생에 유의



## ■ 의료정보 빅데이터 분석모델 개발

- 본 시범사업에서는 건강보험심사평가원의 대용량 보험청구데이터를 분석하여 유의질병 빅데이터 모델을 개발
  - 분석방법으로 데이터마이닝분야에 저명한 기법인 유사그룹 기반 데이터 선별 기법 (Collaborative Filtering)을 사용하였고, 주상병-부상병의 연결성을 정렬하여 기계학습(Machine Learning)에서 사용할 데이터 셋을 생성
  - 아래의 그림은 추출한 데이터로 생성한 연관질병 그래프이며 해당 그래프의 연결성을 의사결정트리의 기준으로 분류하여 사용자에게 유의질병을 선별 제공

## [그림] 연관질병 데이터 그래프



## 효과 및 향후 적용 확대 방안

- 보건의료 빅 데이터를 기반으로 국민건강 증진을 위한 공익 서비스 제공
  - 의료정보의 공개·개방·활용 확대를 통한 맞춤형 서비스 제공 및 정보이용 활성화를 통해 어렵고 복잡하게 만들어진 데이터를 사용자가 보다 쉽게 접할 수 있도록 함으로써 국민건강 증진에 기여함
- 정부 3.0 및 창조경제에 부합하는 새로운 부가가치 및 일자리 창출
  - 데이터 공개 및 Open API 방식으로 빅데이터 분석결과를 외부에서 참고 가능하도록 개방함으로써 신규 비즈니스 창업 및 일자리 창출에 기여함

IV

## 제조





## 14. GE, ‘지능형 항공 운영’ 서비스



항공기의 부품 및 시스템에 장착된 센서를 통해 발생된 데이터를 수집하고 분석, 지능형 항공기 정비를 실현

### 추진 목적 및 배경

- 항공 운영 시 부품 정비 문제로 인한 항공 지연 및 결항이 종종 발생하며 이로 인한 다양한 문제가 추가로 발생
- 항공 지연으로 인한 추가 비용 발생 문제
  - 비행 지연으로 항공사가 지출하는 추가 비용은 연간 약 400억 달러로 추산
  - 비행 지연 중 10%는 예상치 못한 항공기 정비 문제와 관련되어 있으며, 이에 따라 항공기 정비 문제를 사전에 예측하여 해결하고자 하는 필요 발생
- 항공 지연으로 인해 발생하는 승객들의 불편 및 온실가스 배출증가 문제
  - 항공 지연 시 왜 운항이 지연되는지, 언제까지 기다려야 하는지 알 수 없어 승객들의 불안감과 불만으로 인한 고객 항의 발생
  - 이착륙 대기 중인 항공기가 공항 주변에서 선회하는 등 항공교통 정체가 발생하게 되면서 온실가스 배출되는 환경 문제 발생. 특히 고공에서 배출되는 온실가스는 기후변화 등 환경에 끼치는 피해가 타 산업에 비해 큼
- 기계의 정비와 유지보수의 경우 규칙적인 스케줄에 따라 진행되기 때문에 엔지니어들의 불필요한 노동력 발생되는 등 비효율성 증가

- 항공 운항 산업 확대에 따라 예측하지 못한 항공 지연에 따른 금전적 피해도 커지는 등 여러 방면의 문제점 및 비효율성이 제기됨에 따라 문제 해결 필요성 발생

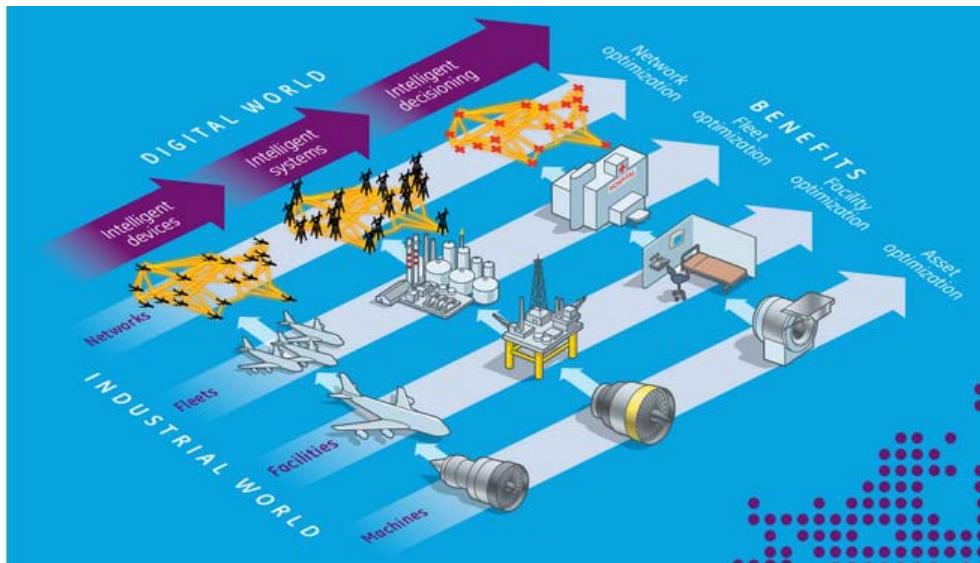
## 추진 내용

- 항공기 부품 및 시스템을 위한 센서 데이터를 분석하고 예측하기 위한 ‘지능형 운영 (Intelligent Operations)’ 서비스 도입
- GE항공(GE Aviation)은 지능형 항공 운영 등을 포함한 산업 인터넷<sup>8)</sup> 기술 개발을 위해 컨설팅 기업 액센츄어와의 합작사인 지능형 운영서비스 회사 탈레리스(Taleris)사 설립
- 다수의 항공기 부품과 부속품 및 시스템의 모든 센서에서 확보된 빅데이터를 모니터링 하여 항공기 정비 문제를 사전에 진단·예측해 지연 출발과 항공편 취소를 사전 예방하고 항공기 정비와 비행 운영을 최적화할 수 있는 탈레리스 지능형 운영(Intelligent Operations) 시스템 개발

---

8) 산업인터넷: 사물인터넷 기술을 철도, 항공, 발전소 등 산업 분야에 적용한 것으로 제품 진단 소프트 웨어와 첨단 분석 솔루션을 결합해 기계와 기계, 기계와 사람, 기계와 비즈니스 운영을 서로 연결시켜 기존 설비나 운영체계를 최적화하는 것을 말함

[그림] ‘지능형 운영(Intelligent Operations)’ 서비스 개념도



자료: GE imagination at work, 2012

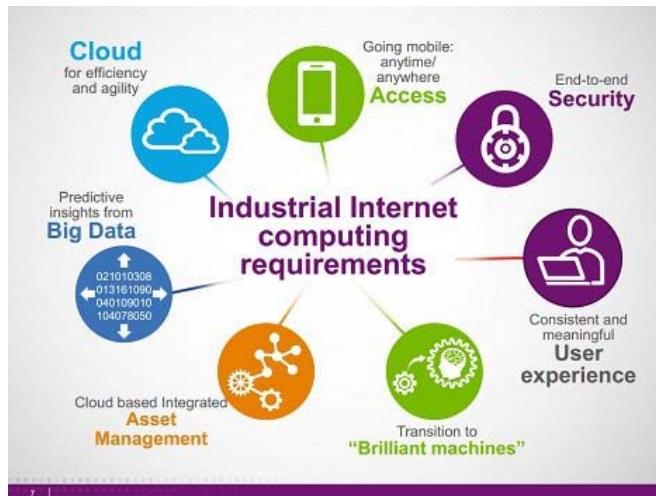
### ※ 지능형 운영 (지능형 항공 운영)

각종 부품과 시스템에 센서가 장착된 ‘지능형 기기(Intelligent Device)’들은 각각의 장비에서 발생하는 데이터를 수집하고 모니터링하는데 이들 지능형 기기들이 네트워크를 통해 서로 연결되어 있는 것을 ‘지능형 시스템(Intelligent System)’이라고 한다. 지능형 시스템은 기기 간의 네트워크 연결 뿐 아니라 수집된 데이터들을 종합하고 분석함으로써 기기 정비 문제를 사전에 진단하고 예측할 수 있다.

항공 산업의 경우 항공기 등이 장비 및 부품 복구 정보는 물론이고, 항공기 정비에 최적화된 시간과 운영인력 배치 등의 정비 계획, 그리고 비행경로 및 운항 스케줄 계획까지 사전에 세우는 등 다양한 방식으로 적용할 수 있다. 이렇게 지능형 시스템을 기반으로 하는 비즈니스 의사결정을 하는 것은 ‘지능형 의사결정 (Intelligent Decision Making)’, 기업 운영을 하는 것을 ‘지능형 운영(Intelligent Operation)’이라고 한다.

- GE 지능형 운영 및 ProDAPS(Probabilistic Diagnostic and Prognostic System, 확률적 진단 및 예측 시스템) 분석 기술을 활용하여 발생 가능한 문제를 미리 예측하여 빠른 의사 결정 지원
  - 항공기 부품으로부터 발생한 데이터를 IoT(Internet of Things) 네트워크를 통해 수집하며, GE 고유의 알고리즘을 이용해 수집된 자료를 실시간으로 모니터링
  - 항공 기계들이 스스로 데이터를 분석하고 공유할 수 있게 하여 관리자에게 의미 있고 유용한 정보를 실시간으로 제공, 발생 가능한 문제의 실시간 파악 및 항공 정비 및 자연 문제를 사전 진단하는 효과적인 운영을 지원
  - 이를 통해 운행 안전 관리, 운항 거리 축소를 통한 연료 절감 관리 등 항공 운영상의 다양한 프로세스 효율을 개선
- 연구개발의 단계를 넘어 전 세계 항공 및 화물운송 기업들에게 서비스 상용화를 계획하는 등 GE 외부로의 서비스 확산을 위해 노력
  - 항공기 성능 데이터, 고장예측 기술, 복구 및 계획 등을 활용해 항공기 효율성 증대 서비스를 전 세계 항공 및 화물 운송 기업들에게 제공 예정
  - 이를 위해 전문 인력 투입 및 연구개발 등 총 10억 달러 투자를 진행
- GE는 산업인터넷(Industrial internet) 전략을 수립하여 기계와 기계, 기계와 사람, 기계와 비즈니스 운영을 서로 연결시켜 기존 설비나 운영체계의 최적화를 지속적으로 추진(GE 2012)

[그림] 산업인터넷 추진방향



## 효과 및 향후 적용 확대 방안

- 지능형 시스템 기반의 ‘지능형 항공 운영(Intelligent Operation)’ 실현으로 운영 효율성 극대화
  - 예측하지 못한 항공 정비로 인해 발생하는 비행 지연을 사전에 예방하여 항공기 운영 효율성의 증가, 정시 운영의 증가, 정비 효율의 증가, 유지보수 비용의 감소를 실현
- 항공 운영 시 발생할 수 있는 사건사고 발생 최소화로 고객만족도 증가와 산업 효율성 제고
  - 항공기 지연, 취소의 최소화로 운영 효율성의 증대는 물론 고객만족도를 향상, 항공산업 전반의 효율성까지 증대

## 15. 볼보, 운행 정보 활용한 자동차 안전 실현



대규모의 자동차 운행 데이터를 기반으로 차량 결함을 예측하고 차량보상의 정확도를 제고하며 고객관리 강화와 기업 경쟁력 향상 실현

### 추진 목적 및 배경

- 안전 운행 실현을 위한 운전 습관, 이동 경로, 검색 정보 등의 차량 빅데이터 수집 및 활용 필요성 증가

### ■ 볼보는 핵심가치인 ‘안전’을 위한 지속적 노력 진행

- SIPS(Side Impact Protection System, 측면 충격 보호 시스템), 후방 어린이 안전시트, 사이드 에어백, 3점 안전벨트, 충돌 경고 및 자동브레이크 등을 개발하는 등 안전을 위한 기술개발 투자를 적극적으로 진행
- 차량 부품과 시스템이 점점 늘어나고 복잡해지며 원인을 파악하기 어려운 다양한 안전 문제가 발생되고 있으며 특히 최근 발생이 잦은 자동차 결함에 따른 사고 및 리콜로 인한 자동차 기업 피해가 증가
- 이를 줄이기 위해 ‘실제사고 연구 - 안전요구 충족 - 제품개발 반영 - 차량 시험 - 생산’의 개발 순환주기를 적용 중이며 더욱 면밀히 문제점을 파악하기 위한 빅데이터 활용 필요

### ■ 빅데이터를 통한 볼보의 ‘안전 철학’ 강화

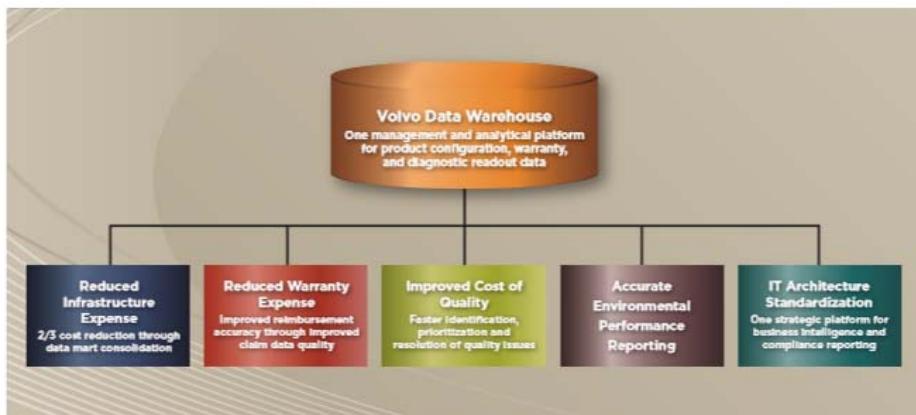
- 빅데이터를 토대로 자동차의 결함을 발견하고 차량의 문제를 파악하여 볼보 자동차의 ‘안전 철학’을 더욱 강화하고자 함
- 실제 발생된 주행 데이터를 통한 분석이 이루어지기 때문에 문제를 더욱 정확하고 빠르게 파악 가능

## 추진 내용

- 많은 양의 데이터를 처리하기 위해 데이터 웨어하우스를 개선 및 빅데이터 수집 등 빅데이터 플랫폼 강화 작업 착수
- 볼보의 초기 빅데이터 사업은 레거시 시스템을 새로운 데이터 웨어하우스로 변경하는 빅데이터 플랫폼 개선 작업의 일환으로 시작
  - 대용량, 실시간 데이터의 효과적 수집, 적재, 활용을 위해 2007년부터 새로운 2-node 테라데이터 5450 웨어하우스를 사용하기 시작
  - 차량 및 하드웨어 사양, 내장 소프트웨어 사양, 차량진단 데이터, 데이터 액세스 및 분석 표준 보고서와 특별 분석을 통해 발생하는 데이터를 수집하기 시작하였으며 수집 된 데이터를 사내에서 개발한 BI<sup>9)</sup>로 구현
- 새로운 빅데이터 플랫폼 기반 마련으로 전산 운영비용 감소 및 시간 절약 효과
  - 새로운 빅데이터 웨어하우스는 처리 할 수 있는 데이터의 양이 364GB에서 1.7TB까지 증가되었으며 데이터 관리를 위한 비용의 절감 효과 발생
  - 또한 차량 주행거리 측정의 경우 기존 2시간이 걸렸다면 5분으로 크게 감축하였고 진단오류코드의 보고서 분석 소요 시간이 2주에서 15분으로 감소하는 등 프로세스 처리 시간이 단축되는 성과를 보이며 빅데이터 분석 기반을 마련

9) BI(Business Intelligence): 기업전략 수립에 필요한 데이터를 수집하고 이 데이터를 이용하여 적절한 의사결정을 내리는데 도움이 되는 일련의 소프트웨어제품군을 의미 [네이버 지식백과]

[그림] 볼보 데이터 웨어하우스 개념도



자료: Teradata(2012. 5)

- 자동차에 센서를 탑재하여 운행 시 발생하는 빅데이터를 적극적으로 수집하고 이 과정에서 나타나는 문제코드들을 수집
  - 엔진, 변속기, 브레이크, 자동 주행 속도 유지 장치, 온도 조절, 승객좌석, 계기판 등 차량 한 대마다 1년 간 발생되는 약 100~150KB의 데이터를 전부 축적
  - 축적된 데이터를 감시하고 진단하며 이러한 진단 과정에 문제가 생길 경우 시스템 진단 문제 코드(DTC: Diagnostic Trouble Code)를 차량엔진(ECU: Engine Control Units)에 저장하여 예측 기반 마련
- 수집된 차량 데이터를 통한 불량 파악, 차량 보상 정확도 제고, 고객 요구사항 파악 등 다양한 비즈니스에 활용
- 운행 데이터 분석을 통한 효과적인 불량 확인 가능
  - 기존 구조로는 새로운 모델을 출시할 경우 약 5만대를 생산해야만 초기 불량을 알아낼 수 있었음

- 빅데이터 도입 후 차량에 센서를 부착해 얻은 데이터를 분석함으로써 약 1,800~2,000대를 생산 할 때 불량을 잡아낼 수 있도록 개선

#### ■ 진단 판독 데이터를 분석함으로써 보상의 정확도를 증가시킴

- 기존에는 차량 보상 진행 시 같은 차량에 대한 고객 제출 정보와 실제 진단 판독 데이터 간에 주행거리에 대한 격차가 발생하여 보상청구 정보에 대한 불확실성이 존재하였음
- 현재 주행데이터를 포함한 모든 진단 판독 데이터를 차량의 ECU(Engine Control Units)로부터 직접 받아 판독하여 문제를 해결하므로 보상의 정확도가 증가

#### ■ 빅데이터를 활용하여 부가가치를 창출하는 개체는 초기 기업 생산/연구 부서에서 마케팅 부서, 판매 부서 등 전사적으로 확대

#### ■ 볼보는 정확한 정보를 모아 적절히 활용하는 것에 포커스를 두고 있으며 이를 통해 새로운 프로세스, 새로운 비즈니스 방식을 개척함

- 빅데이터를 통해 소비자들의 정확한 니즈를 파악하는 것이 가능해졌으며 데이터를 통한 고객 만족도 분석이 가능
- 기존 생산 부서의 결함 발생 최소화와 물류 최적화 뿐 아니라 인력 수급 계획, 직원 만족 평가는 인사와 재무 리스크 모델링 등 기업 내부 운영에도 빅데이터 활용
- 또한 고객 분석을 통한 웹사이트 최적화나 영업기회 창출 등 마케팅 분야, 비즈니스 기획 및 포트폴리오 최적화 등 제품 계획에도 활용하는 등 데이터 기반의 기업 운영이 전사적으로 실현됨

## 효과 및 향후 적용 확대 방안

### ■ 빅데이터에 기반을 둔 기업 의사 결정으로 시장 내 경쟁우위 확보

- 자동차와 관련된 정보들을 축적하여 생산 및 운영 비용절감, 자동차 품질 개선, 기업 이윤 극대화는 물론 친환경 경영까지 데이터 기반 가치창출을 실현 중
- 플랫폼 기반을 구축하는 것에서부터 현 업무 적용까지 20여 년 동안 빅데이터 도입을 위한 다양한 시도가 진행되어 왔으며 향후 현재데이터를 기반으로 한 미래 예측 분야에 주력 예정

## 16. 캐터필러, 직원 및 기기 데이터 분석을 통한 제조 생산성 향상



임직원 및 기기데이터 다양한 데이터를 분석하여 장비 수명 연장과 안전사고 예방을 수행하고 생산성 향상에 활용

### 추진 목적 및 배경

- 회사 성과를 향상시킬 수 있는 주요 요인을 파악하고 연구 착수
- 광범위하게 퍼진 회사의 딜러 네트워크가 회사의 성공을 위한 주요한 역할을 담당한다고 생각하였으며 이들을 분석하고자 하는 요구사항 발생
  - 캐터필러는 세계적인 건설 제조사로 전 세계에 걸쳐 대규모의 딜러 네트워크 (판매중계 지점)를 구축 구축하여 약 13만 명의 직원을 보유
  - 이러한 판매 중계점의 딜러는 장비의 수명을 늘리거나 효율적으로 사용하도록 하는데 도움을 주며, 고객과도 긴밀히 관계를 유지하는 등 주요한 역할 담당
- 2년 동안 미국 내 5개의 캐터필러 판매중계 지점과 그들이 관리하는 총 57개의 대리점을 대상으로 직원몰입도 수준과 비즈니스 성과에 관한 상관성 연구를 시행
  - 6가지의 주요 비즈니스 메트릭스를 선정하여 이와 관련된 직원몰입도와 생산성 간의 관계를 확인하고 이해하기 위한 작업 착수
  - 가장 좋은 성과를 내는 대리점을 분석하고 그것으로부터 교훈을 얻기 위해 포커스 그룹을 조직

## 추진 내용

- 임직원 조사 데이터, 기기에서 발생한 데이터 등 다양한 데이터에 대한 활용도를 높이는 데 노력
- 직원 몰입도를 성과와 연결시키기 위한 기초 분석 작업 진행
  - 캐터필러의 인력 담당 부서는 글로벌하게 구축된 딜러 네트워크에서 직원 몰입도와 비즈니스 성과와의 관계에 대한 궁금증 보유
  - 두 변수와의 상관관계를 이해하고, 이를 통해 얻어진 통찰력을 활용하기 위한 계획 차수
  - 이를 통해 판매중계 지점의 개선된 활동 계획 수립을 위해 정보를 모음
- 비즈니스 수익 향상과 조직 차원의 성과 향상을 위해 인력 자원의 의견을 기초 자료로 활용하는 ‘피플 이니셔티브’를 시행
  - 캐터필러 인력 담당 부서는 딜러(임직원)들에게 공통적으로 적용되는 성과 메트릭스를 도출하기 위해 그들을 밀접하게 관찰하고 자료를 수집
  - 성과 메트릭스는 비즈니스 성과에 영향을 주는 주요 요소들로 건설 제조업이라는 산업의 특성을 잘 반영하기 위한 추출 작업 시작
  - 이에 따라 안전, 품질(재작업), 소모, 고객 로열티, 기술자 생산성, 수익과 수익 목표간 상관성의 6가지의 비즈니스 메트릭스를 구성
  - 주요 비즈니스 메트릭스에 따라 비즈니스 수익 향상과 조직 차원의 성과 향상이 어떠한 영향관계를 갖는지 분석



## ■ 캐터필러의 상관도 분석을 통해 직원 몰입도가 비즈니스 성과에 직접적인 효과가 있다는 것이 밝혀짐

- 연구는 높은 몰입도를 보인 딜러를 보유한 대리점의 6가지 주요 비즈니스 메트릭스(안전, 품질(재작업), 소모, 고객 로열티, 기술자 생산성, 수익과 수익 목표간 관계)를 측정한 결과 훨씬 좋은 결과를 보인다는 것을 발견
- 높은 직원 몰입도를 보유한 지점은 분기 목표수익 달성을 비율이 그렇지 않은 대리점에 비해 40% 높았고, 고객 로열티는 5.3% 높았으며, 기술자 생산성은 4.5% 높다고 분석
- 또한 이직률이 상당히 낮고, 재작업에 쓰이는 기술자 비용도 낮으며, 사고율도 적다고 나타남

## ■ 조직의 최대 자산은 직원이라는 믿음 하에 매년 캐터필러 임직원 의견 조사(EOS, Employee Opinion Survey)를 시행하기로 계획

## ■ 직원 데이터 분석 뿐 아니라 각 장비에 축적된 데이터를 통해 장비 수명 연장 및 건설 현장의 안전사고를 예방하기 위한 분석 작업 진행

- 각 장비마다 GPS와 센서, 데이터 관리 소프트웨어를 설치해 기계의 현재 위치, 가동 시간, 가동 상황, 연로 잔량 등의 데이터를 실시간으로 수집
- 이를 통해 장비가 과열되고 있는지, 부속 장치에 이상이 발생하고 있는지 등의 여부를 실시간으로 모니터링 하고 작업자에게 통보
- 관리자는 사무실 모니터를 통해 장비 위치 및 작동 상태 유무, 연료 소비 현황, 위험 신호 파악 등 전반적인 내부 관리 가능
- 장비의 고장을 미리 탐지하여 대응할 수 있기 때문에 고장으로 인한 작업 중단을 미연에 방지하는 것이 가능하고, 예측을 통한 신속한 의사 결정이 가능해짐

## 효과 및 향후 적용 확대 방안

- 회사 제1의 자산인 직원의 성과 조사를 통해 회사의 생산성과 관련한 주요 지표들을 분석하여 개선하고, 그러한 결과를 변화와 성공의 기반으로 활용하기 위해 공유
  - 이러한 작업을 통해 직원의 작업 능률 향상, 근속년수 향상, 작업의 질적 수준 향상을 위한 요인들이 무엇인지 파악 가능
  - 또한 비즈니스 지표별 개선이 필요한 부분과 잘 되어가고 있는 부분을 평가하고 향후 변화를 미리 예측 가능
  - 조직차원의 인력 투자 방법과 투자 대상에 대한 의사결정에 가이드 역할을 해주어 조직 이익 극대화에 기여
- 회사의 주요 자산 중 하나인 장비의 실시간 분석은 부서 의사 결정이나 신속한 고객 대응, 관리자의 사업 방향 설정 등 다양하게 활용 가능
  - 글로벌 공급 소스를 추적하는 것이 가능해지고, 변경사항에 대해 빠르게 반응할 수 있게 되는 등 공급망 관리 개선
  - 데이터를 활용하여 다양한 시나리오를 구성하고 이에 대한 예측을 도출 가능하기 때문에 사고에 즉각적으로 대응 가능
  - 회사 성과를 시각적으로 파악 및 분석할 수 있게 됨에 따라 수익성과 ROI를 보다 쉽게 이해하여 빠른 의사결정 가능
- 빅데이터 활용을 통한 기업 이미지 쇄신 효과
  - 생산성 향상, 고객 로열티 증가 뿐 아니라 전 세계의 수많은 고객들과의 관계를 지속적으로 유지하며, 기업 특유의 이미지를 각인시키는데 도움

## 17. 한국남동발전, 발전설비 운영효율 극대화



시스템 및 연료 설비로부터 발생되는 데이터를 수집·분석하여  
연료비 절감 및 안정적인 설비 운영 추진

### 추진 목적 및 배경

- 연료자원 관리 효율성 제고 및 설비 장애 발생 피해 최소화를 위한 빅데이터 활용 필요성 제기
- 발전원가의 대부분이 연료구매 비용으로 연료자원관리 최적화 필요성 대두
  - 삼천포와 영흥에서 1일 석탄사용량은 65,000톤 수준이며, 이는 25톤 덤프트럭 2,600대, 구매비용으로 환산할 경우 약 70억 수준으로 추정
  - 발전 원가의 76% 이상이 연료구매 비용으로 연료자원관리 최적화의 필요 발생
- 저열량탄 사용으로 열효율 향상시킬 수 있는 방안 필요
  - 발전원가의 70% 이상을 차지하는 연료비 절감을 위해 저열량탄 사용이 지속적으로 증가되는 상황에서 설비 부하를 최소화하며 열효율을 획기적으로 향상 시킬 수 있는 방안모색 필요 발생
- 핵심설비의 장애 발생은 엄청난 경제적 손실로 안정적인 전력생산을 위한 장애 발생 최소화
  - 발전기 1개 호기를 구성하는 설비는 보일러, 터비, 발전기, 통풍장치 계통 등 총 20만개의 Unit으로 구성되며 발전소 운전에 직접적인 영향을 미치는 핵심설비 또한 1만개 이상으로 구성

- 이러한 핵심설비의 장애 발생은 엄청난 경제적 손실을 가져오기 때문에 발전 설비를 안정적, 효율적으로 관리 할 필요 발생

## ■ 발전설비 고장 시 장애 복구까지 많은 시간 소요

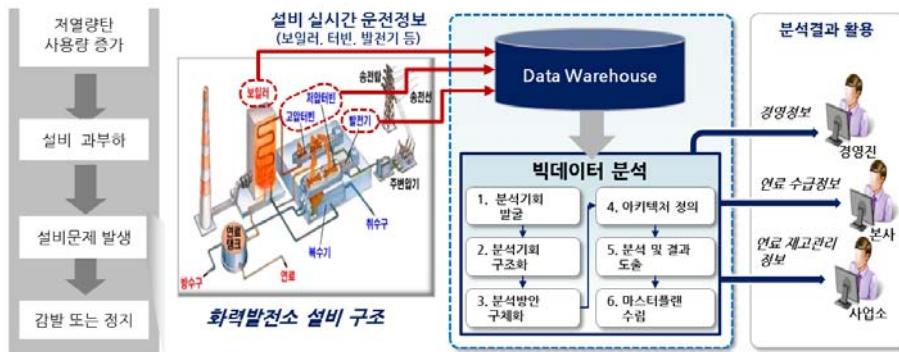
- 발전설비 고장이 발생하게 되면 재가동 등의 장애 복구까지는 많은 시간이 필요 하여, 사전에 고장을 예측함으로써 복구비용을 절감하고자 하는 요구사항 발생

## 추진 내용

### ■ 연료효율화를 위한 빅데이터 분석 체계 수립

- 저열량탄 사용증가에 따른 설비 과부하, 출력 감발 등의 설비문제를 예방, 예측하기 위해 시스템 및 연료설비로부터 발생되는 데이터를 수집 및 분석하고, 분석된 결과를 기반으로 연료비 절감과 안정적인 설비운영 추진

[그림] 한국남동발전(주)의 빅데이터 분석 구조

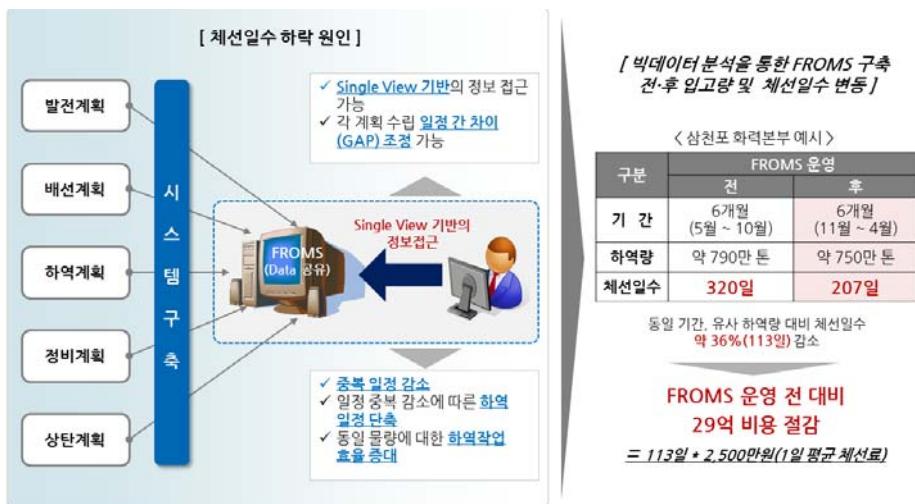


자료: 한국남동발전, 2014

## ▣ 각 단계에서 필요한 데이터를 통합/공유하여 프로세스 단축

- 전기를 생산해내기 위한 5가지 계획(발전계획, 배선계획, 하역계획, 정비계획, 상단계획)을 단일관점(single View) 기반으로 시스템화 한 FROMS<sup>10)</sup>를 통해 각 단계에서 필요한 데이터를 통합/공유함으로써 프로세스가 단축
- 한국남동발전(주)의 삼천포의 하역작업이 FROMS 구축전보다 113일 정도 단축, 이를 통해 당사의 발전비용 절감 및 생산성 향상에 공헌

[그림] 한국남동발전(주)의 FROMS를 통한 프로세스 개선



자료: 한국남동발전, 2014

## ▣ 발전설비 운영 최적화를 위한 분석

- 보일러 튜브의 국부과열 진단을 위해 출력, 연료성상, 운전변수 등 상관분석 및 추세분석을 통해 운전 제한치 도달 예측을 함으로써 적정한 정비시기를 결정 가능

10) FROMS : Fuel Resource Optimization Management System, 석탄최적관리시스템

- 다양한 내·외부 데이터를 수집 및 분석하여 설비 운전상태 여부에 따라 패턴화하고, 실시간 운전이 이상 징후를 보일 경우, 기정의된 유사 패턴과 비교하여 고장발생 알림 및 적정한 정비시기를 결정 가능

## ■ RBM/RCM 등 선진정비기법을 도입

- 데이터 분석을 통한 고장 및 주기적 계획예방정비에서 설비 상태기반의 예측정비 체계로 전환

※ TBM : Time Based Maintenance(시간기준 계획예방정비)

CBM : Condition Based Maintenance(상태기반 정비)

RBM : Risk Based Maintenance(위험도기반 정비)

RCM : Reliability Centered Maintenance(신뢰도기반 정비)

GENi : Growth Engine for New Integration(발전설비관리시스템)

[그림] 한국남동발전(주)의 예방정비 및 예측정비



자료: 한국남동발전, 2014

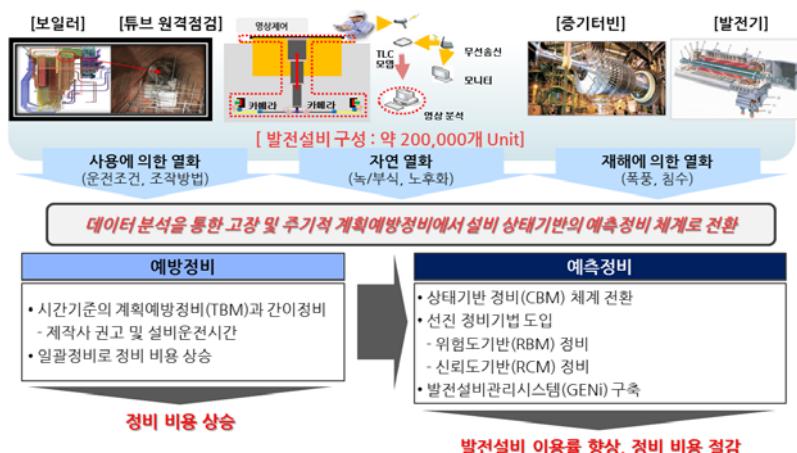
■ 발전 설비에 대한 정보를 계획정비에서 상태기반 정비체제로 전환하기 위한 빅데이터 활용

- 전기 생산을 위한 발전 및 연료 설비에서 발생하는 센서 데이터를 이력정보로 구축하고, 해당정보와 정비계획 및 점검결과 간의 상관관계분석을 통하여 장애 패턴 및 빈도를 이용한 상태예측과 장애대응방안 수립체계 구축을 추진

■ 데이터 기반의 최적 혼란비율 선정 업무 지원을 위한 빅데이터 활용

- 기준에는 혼란관리시스템 또는 운영 및 예측관리시스템에서 제공되는 복수의 혼란 시나리오를 담당자의 경험과 지식에 기반하여 결정하였으므로 업무 순환 배치 등으로 담당자가 교체되는 경우 담당자의 개인적인 능력에 따라 최적 시나리오 선정 결과에 차이 발생
- 담당자들이 선정한 시나리오를 바탕으로 실행한 발전운영결과에 대한 유사성, 빈도 및 패턴 분석 등 빅데이터 요소를 가미하여, 유사한 조건에서 실행한 과거의 운영이력 분석데이터를 근거로 복수의 시나리오 중 최적안을 선택할 수 있는 활용체계 구축

[그림] 경험과 지식 기반의 최적 혼란 비율, 사전 장애 예측 및 신속 대응 개념도



자료: 한국남동발전, 2014

## 효과 및 향후 적용 확대 방안

### ■ 발전설비 운영효율 극대화를 통해 생산성 향상 도모

- 단위시스템 연계와 단일관점 기반의 통합시스템 운영 및 지속적인 고도화 작업을 통해 데이터 기반의 일관된 정보를 제공하고 공유함으로써 의사결정의 정확성과 신속성을 제고하여 업무 리드타임 단축과 이에 따른 생산성 향상 도모

### ■ 분석데이터 기반의 명확한 의사결정 실현

- 빅데이터 분석체계 수립을 통해 개인의 직관이 아닌 분석데이터 기반의 명확한 의사결정이 가능해짐으로써 프로세스 개선과 운용비용 절감 등을 통한 업무 효율성 및 생산성 증대와 타사 대비 경쟁력 강화

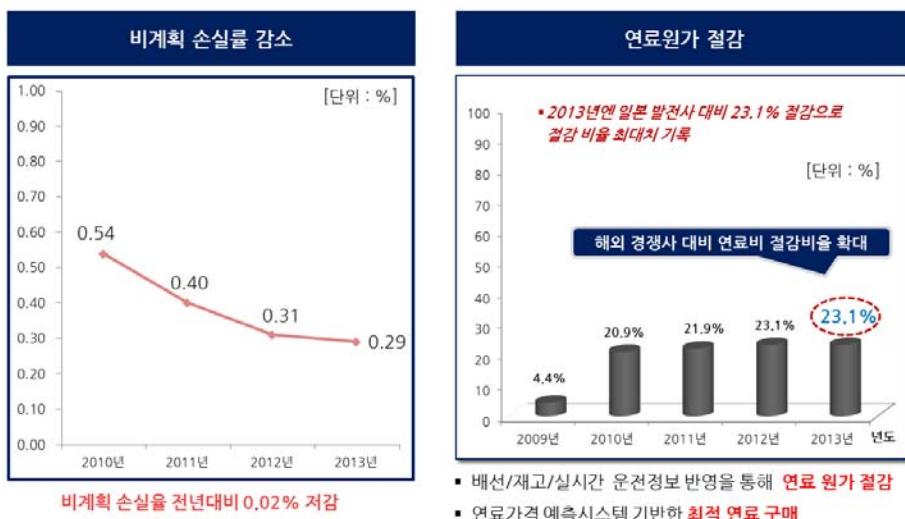
### ■ 빅데이터 분석의 업무 프로세스 내재화

- 업무 프로세스 수행품질의 지능화를 위해 빅데이터 분석을 발전설비 운영과 연료 운영 업무 프로세스에 내재화하여 실시간 의사결정에 따른 작업 수행과 연속적 피드백 유도

### ■ 발전설비관리시스템의 효율적인 운영을 통해 경영혁신 성과

- 발전설비관리시스템의 효율적인 운영을 통해 ‘비계획 손실률 감소’, ‘설비 이용률 증가’, ‘평균 무고장 시간 증가’ 등 혁신 성과 도출

[그림] 한국남동발전(주)의 추진성과도표



자료: 한국남동발전, 2014

## 18. 자동차 부품기업 공동활용 빅데이터 플랫폼



데이터 분석 기반의 제품 품질향상을 위해 자동차 부품 제조사가 공동 활용 할 수 있는 빅데이터 플랫폼 구축 운영

### 추진 목적 및 배경

#### ■ 사업 추진의 배경

- 한국의 자동차 산업은 2013년도 전체 452만대 생산하여 세계 점유율 5.2%로 9년 연속 세계5위로 산업의 10% 이상 점유하여 가장 큰 제조 분야를 차지
- 빠르게 성장하고 있는 중국 자동차 생산량은 2013년 2212만대로 전체 1위를 하였으며 최초 2000만대 돌파
- 빅데이터 서비스 가치에 대한 인식부족으로 자동차산업의 경쟁력을 좌우하는 자동차부품 업체의 공장/품질 정보에 대한 분석 서비스 부재



## ■ 사업 추진의 필요성

- 2012년 한국자동차산업은 유럽 발 경제위기 및 고유가 속에서도 한국차의 품질 및 신뢰도 향상, FTA 발효 확대 등으로 수출 3백17만대, 수출금액(부품 포함) 718억 달러로 사상 최대 수출실적을 기록하였지만, 국내 시장의 경기 불확실성 지속 및 가계부채 증가에 따른 소비심리 위축으로 국내 생산은 3년 만에, 내수판매는 4년 만에 각각 감소세를 나타냄

### ※ 자동차 산업 현황

- \* 생산: 내수부진 및 임단협 기간 중 공급차질로 3년 만에 감소세로 전환
- \* 내수: 경제성장을 하향과 가계부채 부담, 신차효과 약화 등으로 감소세를 보임
- \* 수출: 국산차의 품질 및 브랜드 이미지 상승, 주요국과의 FTA 체결로 인해 경쟁력 상승
- \* 수입: 최근 수입차에 대한 소비자 인식 제고 등으로 국내 수입차 시장규모가 급격히 증가

- 자동차산업의 국민경제적 비중(2만 여개 부품으로 생산되는 전후방 연관효과가 가장 큰 산업)은 제조업 생산의 11.4%, 고용의 10.7% 및 부가가치의 10.6% 차지(2011 통계청 광업제조업통계 조사보고서 참고)
- 2012년 기준 자동차산업의 무역수지는 616.5억불로 반도체, 선박류 등 주력 기간산업과 비교해 볼 때 국가 경제에 이바지 하는 바가 큼
- 연계 산업의 파급효과와 고용효과가 큰 자동차산업의 경쟁력 강화를 위해 강소 자동차 부품기업 육성이 필요하고, 이를 위해 자동차부품들이 공동으로 활용할 수 있는 데이터 분석을 통한 품질 및 생산성 향상 공동 서비스 제공 필요

※ 데이터 분석은 기존 시스템의 정보를 활용하여 추가 투자가 많이 필요치 않고, 공정개선 만으로 품질과 생산성 향상이 가능

## 추진 내용

| 참여기관 | 메타빌드(주), 자동차부품연구원, 솔바테크놀로지

### \* Data 제공업체(자동차 부품기업 K 사)

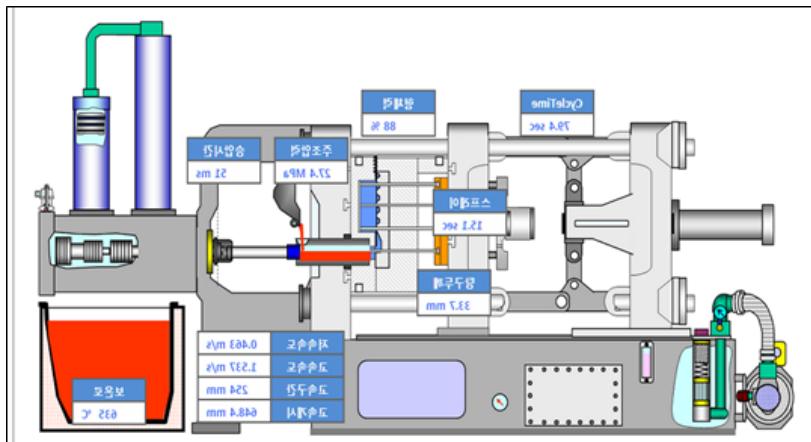
- K사는 3차 기업으로 주조 가공 전문 업체이며 다이캐스팅 주조 가공 전문 기업이다. 칠천만불 수출의 탑을 수상하였으며 동종업계최초 ISO 16949 인증, 중소기업인 대통령상 수상을 한 바 있다.
- 매출 중 내수 및 해외가 54%, 공조, 엔진, 변속기, 조향 부품에서 46%를 차지하고 있으며, 내부 주조 관리 MES(Manufacturing Execution System)가 가동 중에 있다. 전문 가공 공장 시스템을 운영하고 있으며 추가 도입을 계획중에 있다. 본 사업을 위해 품질관련 데이터 제공 및 업무 공조 협약 맺어 데이터를 제공하였다.

| 주요 활용데이터 |

구분	데이터	데이터 규모	보유기관
설비 및 품질데이터	설비 설정값 실측정보	564,946건(540MB) *표본 샘플링 6개월 치	K 사
조건별 생산수량 대 불량수량 데이터	2만건(14MB) *표본 샘플링 6개월 치		
작업장 온습도 정보	24만건(70MB) *표본 샘플링 6개월 치		

- 다이캐스팅 주조공정 장비 설정데이터, 실측데이터 6개월치(MES에서 추출)

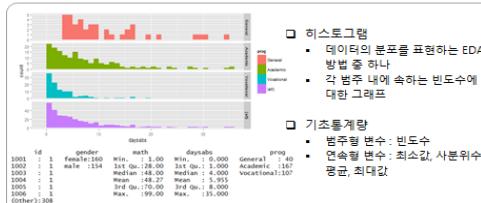
[그림] 분석대상 알루미늄 다이캐스팅 설비



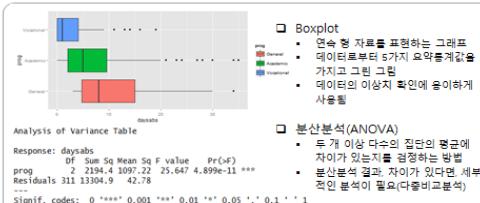
## |분석 내용 및 기법|

### 데이터 분석 기법

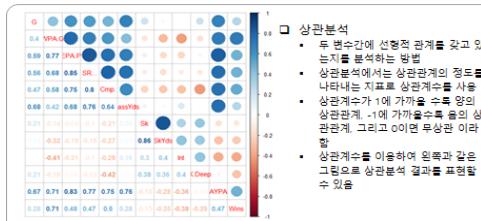
#### EDA: 기초통계량, 히스토그램, Heatmap 등



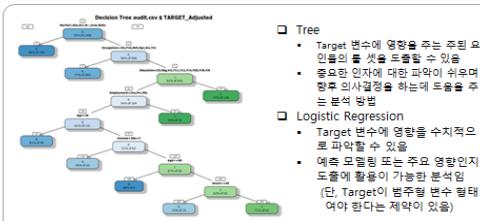
#### 호기 별 차이 분석: T-test, ANOVA, Boxplot 등



#### Parameter 간 분석: 상관분석

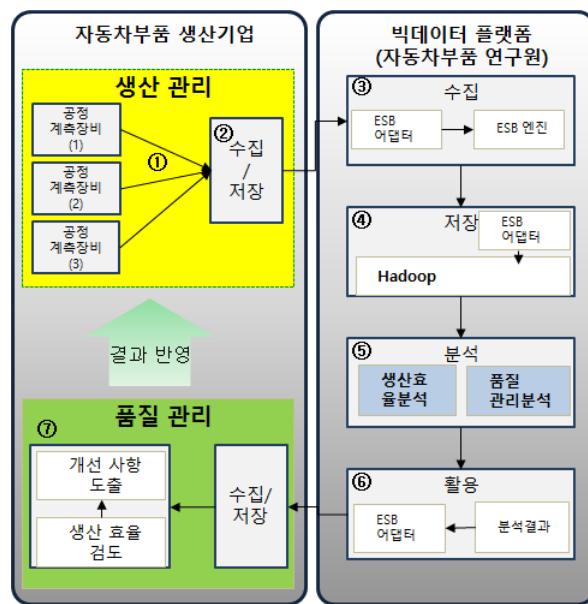


#### 주요 Parameter 분석: Tree



## 데이터 분석과정

[그림] 데이터 분석과정



### ① 데이터 수집

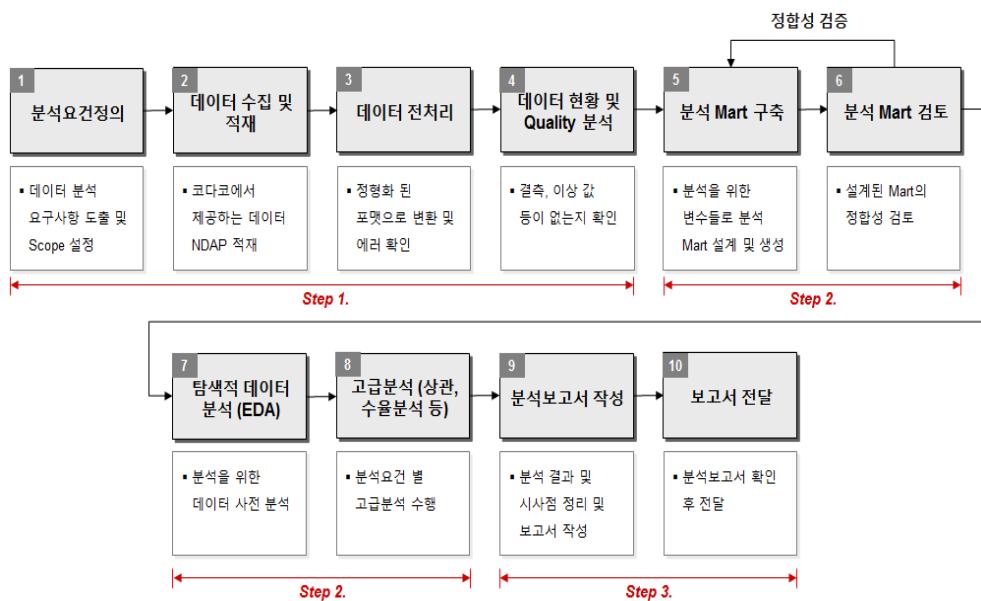
- 분석된 결과 데이터를 해당 부품기업의 전용 웹 사이트에 정보 제공
- 필요한 내부 시스템으로 연계 전송 반영

### ② 데이터 저장

- 자동차부품 공통 분석 플랫폼의 하둡서버에 추출된 데이터 저장

### ③ 데이터 분석

[그림] 상세 데이터 분석 절차



#### ● 탐색적 데이터분석

- 분석 대상으로 선정된 데이터의 탐색적 분석을 통해 다양한 분석 포인트를 도출하고 변수별 기초통계량 파악, 변수 별 패턴 및 분포 확인

#### ● 각 호기 별 차이 분석

- 변수 별 차이 분석: 작업 호기, 프로세스, 작업시간, 작업자 등 조건에 따른 불량률 및 수율차이 분석

[그림] 작업조건별 불량률 분석

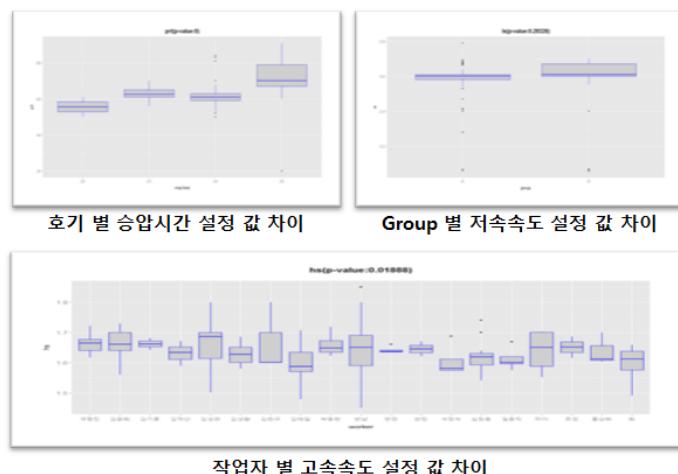


### ● 설비 설정 값 간 상관 분석

- 6개의 설비 설정 값의 현황 분석, 시간 별 패턴 분석
- 주조에 사용되는 설비 설정 값 간의 연관성을 상관분석으로 분석

## [그림] 속성별 불량률 분석

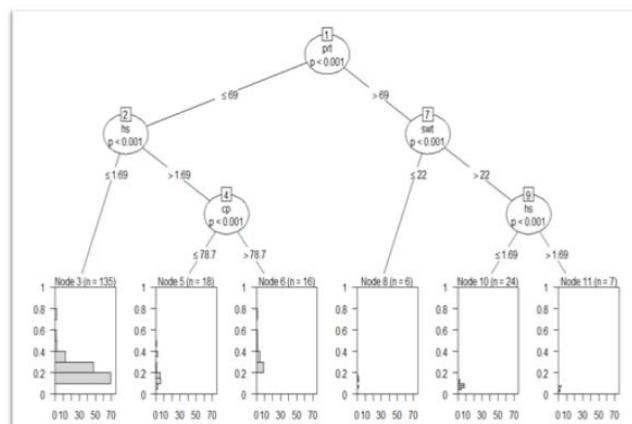
[ 속성 별 Parameter 설정 값에 대한 ANOVA ]



### ● 불량률에 영향을 주는 요인분석

- 통계 모델링을 통하여 불량률에 영향을 주는 요인을 분석

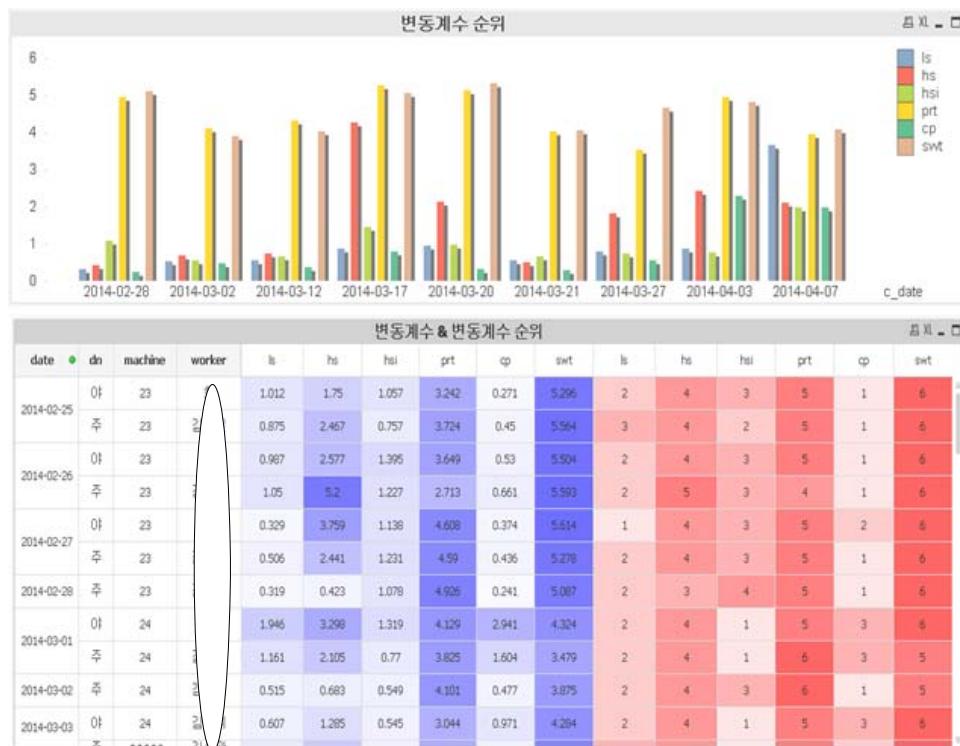
## [그림] 의사결정나무기법을 활용한 요인 분석



## 설비설정값 중 수율에 영향을 주는 주된 요인 분석

- 설비설정값과 실측값 사이의 차이분석, 설비설정값과 실측값의 변동성 분석을 통해 작업자 별로 각 설정 값을 적절히 조절하고 있는지 파악

[그림] 작업자별 변동계수 분석



#### ④ 분석결과 활용

- 분석결과의 현장적용

- 인터뷰 조장의 주조조건 설정 방안을 수평교육 후 결과 확인

[표] 전/후 조건설정치 비교

구 분	생산장비	전체평균							
		생산수량	불량합계	불량률	저속속도	고속속도	고속구간	승압시간	주조압력
1~6월	23,24,25	334.2	34.5	10.3%	0.59	1.64	244.9	63.4	77.9
9~10월	24	405.8	26.7	6.6%	0.61	1.58	232.1	58.3	69.2
차이	*	+71.6	-7.8	-3.7%	+0.02	-0.06	-12.8	-5.1	-8.7

- 전체적인 주조조건의 변경 차이가 나타남
- 주조압력을 기준으로 고속속도부터 변경 적용 후 나머지 조건을 미세하게 변경
- 외형적인 결과를 고려했을 때, 약 3.7%의 불량률이 감소하였음

현장 적용 후의 성과: 불량률 감소에 따른 생산비용 절감 기대

### 주요 분석 결과 및 활용방안

#### |주요 분석 결과|

##### ■ 불량률에 영향을 주는 장비 설정 변수 도출

- 분석 대상으로 결정된 '자동차 변속기 하우징' 생산장비인 알루미늄 다이캐스팅 장비의 생산 수율에 영향을 미치는 변수 도출  
※ 사례) 장비 승압시간 설정값이 69 이하이고, 고속 속도가 169를 초과하며 주조압력이 78.7을 초과하는 경우 0.2 이상의 불량률이 발생

## ■ 동일 공정에 대해 작업자별 운영방법 상이

- 불량률 저하를 위한 장비 운영 매뉴얼 없이 숙련자의 직감에 의존하는 형태로 운영
- 작업자별 변수 설정값이 다르고, 불량률 차이도 심함

[그림] 작업자별 불량률 분석결과

작업자별 상관분석						
	ls	hs	his	prt	cp	swt
ls	1	0.14	-0.893	-0.387	0.179	-0.206
hs	0.14	1	-0.211	-0.004	0.435	0.274
his	-0.893	-0.211	1	0.094	-0.177	-0.013
prt	-0.387	-0.004	0.094	1	0.302	0.088
cp	0.179	0.435	-0.177	0.302	1	-0.392
swt	-0.206	0.274	-0.013	0.088	-0.392	1

작업자별 상관분석						
	ls	hs	his	prt	cp	swt
ls	1	-0.52	-0.155	0.087	-0.436	-0.134
hs	-0.52	1	0.255	0.101	0.109	0.163
his	-0.155	0.255	1	0.088	-0.211	-0.631
prt	0.087	0.101	0.088	1	0.082	-0.123
cp	-0.436	0.109	-0.211	0.082	1	0.254
swt	-0.134	0.163	-0.631	-0.123	0.254	1

[김○○: 평균 불량률- 13%]

작업자별 상관분석						
	ls	hs	his	prt	cp	swt
ls	1	-0.265	-0.686	0.298	-0.408	0.569
hs	-0.265	1	0.536	0.068	0.276	-0.373
his	-0.686	0.536	1	0.093	0.544	-0.642
prt	0.298	0.068	0.093	1	-0.433	0.428
cp	-0.408	0.276	0.544	-0.433	1	-0.869
swt	0.569	-0.373	-0.642	0.428	-0.869	1

[○○○: 평균 불량률- 7%]

[김○○: 평균 불량률- 13%]

작업자별 상관분석						
	ls	hs	his	prt	cp	swt
ls	1	0.112	-0.045	0.104	-0.021	-0.153
hs	0.112	1	0.196	-0.26	-0.227	0.189
his	-0.045	0.196	1	-0.386	0.126	0.147
prt	0.104	-0.26	-0.386	1	-0.214	0.259
cp	-0.021	-0.227	0.126	-0.214	1	-0.302
swt	-0.153	0.189	0.147	0.259	-0.302	1

[○○○: 평균 불량률- 14%]

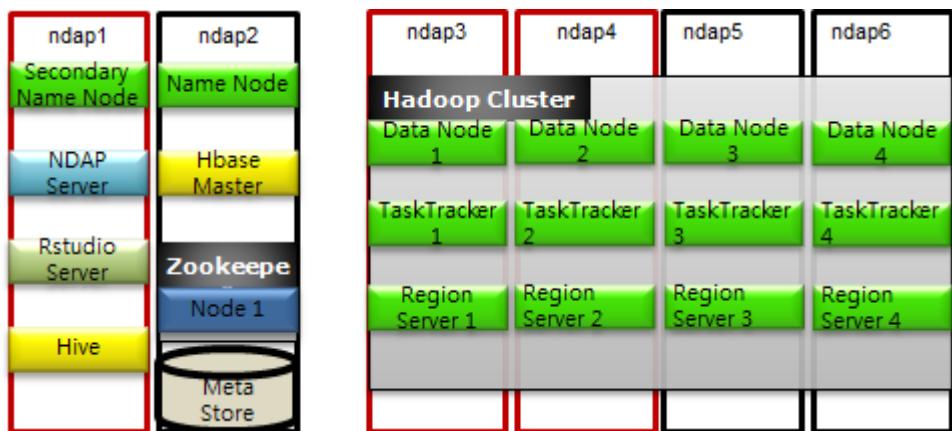
## | 서비스 계획 및 활용방안 |

### ■ 장비 운영 매뉴얼 제작

- 분석 대상을 확대하고, 정밀 분석을 거친 후 장비 운영 매뉴얼 및 가이드 제작에 활용 예정
- ※ 분석결과를 반영하여 장비운영 조건을 실험한 결과 3.7%의 불량률 감소

- 분석 지원 대상 기업 확대 및 자동차 부품기업 공동활용 빅데이터 플랫폼 운영계획 수립·추진

[그림] 자동차 부품기업 공동활용 빅데이터 플랫폼



※ KT NexR - NDAP 시스템(하둡 기반)

## 효과 및 향후 적용 확대 방안

### ■ 기대효과

- 부품 품질 향상은 생산 및 공급 수율을 증가시켜 직접 매출 상승에 기여할 것으로 기대(약 10% 예측)
- 부품데이터 분석 기반의 생산품질 향상(불량률 저하를 통한 수익향상)
- 제조업 빅데이터 활용 우수사례 제시

## ■ 사업 활용 방안

- 자동차 부품 기업 공동 활용 빅데이터 분석 플랫폼은 자동차 부품 연구원에 등록된 중소기업 대상의 공공 인프라로 사용을 원하는 중소 자동차 부품 기업 사용 신청 시 일정 조정 및 승인 및 데이터 분석 업체 리스트 제공 등 모든 행정적 절차를 자동차 부품 연구원이 전담 운영 인력을 배치하여 관리
- 빅데이터 분석 컨설팅 사업을 원하는 기업을 대상으로 현 분석 상황 및 진행 절차, 과정, 리소스 등의 다양한 정보를 구체적으로 제공 창업을 유도 지원
- 센서를 통한 데이터 수집을 연구 기획하여 효율적인 공통 데이터 수집 방안과 체계를 마련하여 제시 하는 과제 개발 기회
- 동중업계 시범 사업자와 협조하여 분석 방법론의 기술적인 교육 기회 부여하고 기술 개발을 통한 추가 확장 서비스 개발 시 플랫폼 시범 활용 서비스 기회 제공

V

# 재난·공공





## 19. 농림수산식품교육문화정보원, 스마트 농정 실현을 위한 빅데이터 서비스 구현



농림축산식품 관련 공공데이터와 민간의 정보자원을 활용한 농수축산물 가격정보 및 소비자 유형별 관심 농식품 추천 서비스 제공

### 추진 목적 및 배경

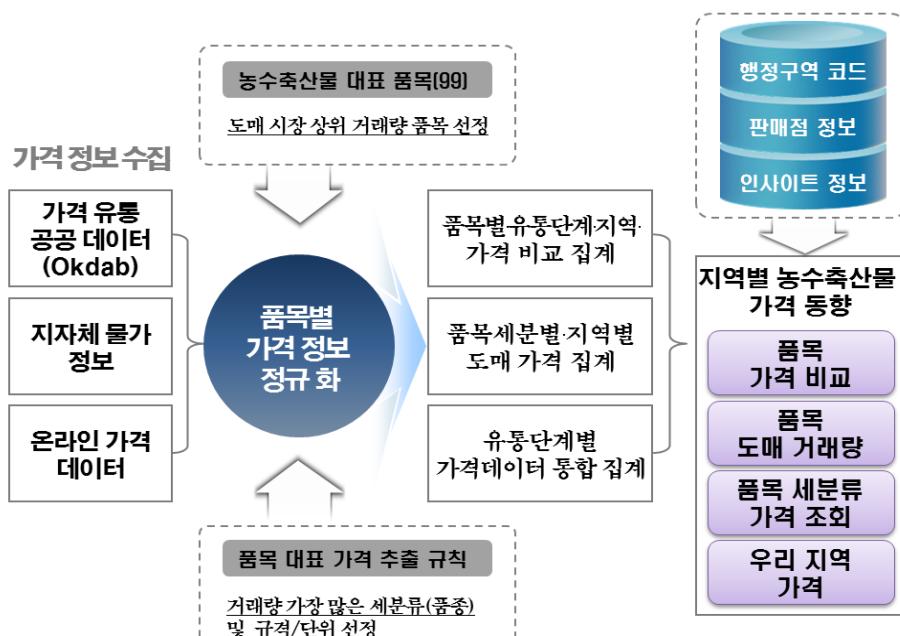
- 농식품 관련 공공데이터를 개방하여 다양한 활용 니즈를 발생시키기 위한 목적
  - 각 분야에서 빅데이터 활용이 확산됨에 따라 국내에서는 수요가 높은 공공 정보를 시작으로 개방에 대한 요구가 확산되었으며, 이를 활용하기 위한 요구사항도 증가
  - 국내 농·식품 공공데이터에 대한 통합서비스 제공이 미흡하고 빅데이터 분석 특성이 반영된 다차원 분석 인프라가 부재한 상황
- 공공데이터를 개방하고 빅데이터 기반 서비스를 제공하여 스마트 농정 구현
  - 이를 통해 공공데이터 활용의 이용 편의성을 확보하고, 빅데이터 활용을 위한 통합 플랫폼을 구축하며, 빅데이터 기반의 다양한 서비스 모델 수립

### 추진 내용

- 빅데이터 기반 마련 및 이를 통한 다양한 분석 서비스 제공
  - 농산물, 수산물, 축산물에 대해 산지(공판장) 가격과 도매(경락)가격, 도소매 가격의 데이터 등 기관별로 분산되어 있는 가격 데이터를 빅데이터 플랫폼 (ASP)을 통해 수집하고 실시간 조회가 가능하도록 구성

- 유통 단계별 가격동향, 도매(경락) 거래량, 지역별 도매 거래량 및 가격 정보 등의 기능을 제공하는 통합 서비스를 구축함
- 트위터, 블로그, 언론, 게시판 등에서 데이터를 수집하여 농식품 대표 품목 추출, 주제 키워드에 대한 버즈량과 연관어 추출, 주요 원문 제공 등 민간 기관의 정보자원을 적극 활용

[그림] 농림수산식품교육문화정보원 빅데이터 서비스 구조 및 제공 기능



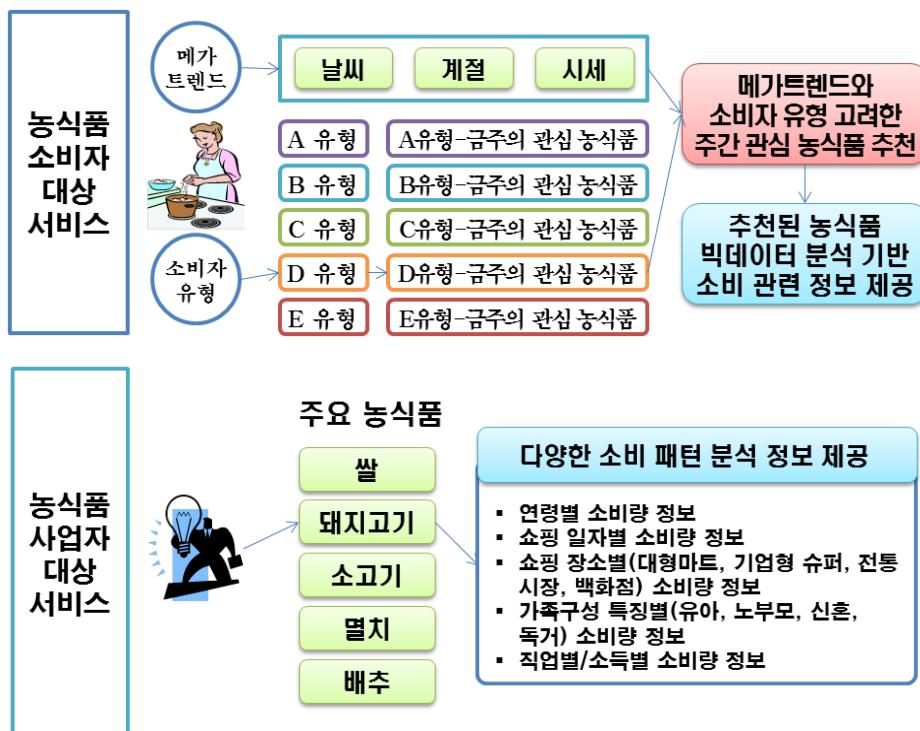
- (1) 가격비교 집계 생성 시 단위 규격(Kg) 당 가격으로 환산
- (2) 산지/도매가격은 대표 가격 추출 규칙관리
- (3) 농수축산물 대표 품목의 월별 주별 가격 정보 제공
- (4) 판매점 정보 DB를 이용한 위치기반 가격 정보 제공
- (5) 수집처별 가격 정보 원본 데이터 제공

자료: 농림수산식품교육문화정보원, 2014

■ 소비 유형별 구매 예측을 분석하는 등 빅데이터 활용·분석 알고리즘 개발을 통해 이용자들에게 인사이트를 제공

- 5가지 식생활 라이프스타일 유형을 도출하였으며 이를 날씨, 계절, 시세 등의 메가트렌드와 함께 고려하여 소비자 유형별 관심 농식품을 추천해주는 서비스 개발
- 소비자 유형별 추천 메뉴를 제시하여 주고 이와 관련한 식자재 내역과 사용자 지역의 가격, 온라인 가격, 연관 키워드 등을 조회할 수 있는 분석 서비스 제공

[그림] 5가지 식생활 라이프스타일 유형에 따른 농식품 추천 서비스 구성도



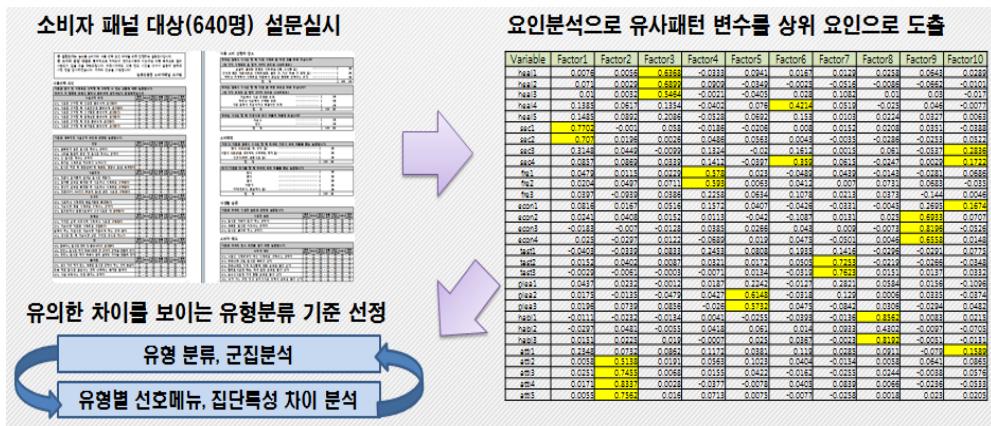
자료: 농림수산식품교육문화정보원, 2014

## \* 참고 : 소비자 유형별 관심 농식품 추천 서비스 알고리즘 개발

### <1> 식생활 라이프 스타일에 따른 5가지 유형 도출

- 농진청의 소비자패널 640명을 대상으로 설문조사 진행
- 요인분석을 통해 유사패턴 변수를 상위 요인으로 도출
- 유의한 차이를 보이는 유형들을 분류 기준으로 선정하여 집단 특성 차이 분석, 유형별 선호메뉴 선정 등 군집분석 시행
- 오피니언 마이닝, 소비자 패널 데이터로부터 유형별 소비패턴 추출이 최대화 될 수 있는 구분으로 최종 결정  $\Rightarrow$  5가지의 식생활 라이프스타일 유형 도출

[그림] 5가지의 식생활 라이프스타일 도출 과정 및 특징



### 5가지 식생활 라이프스타일 유형 도출

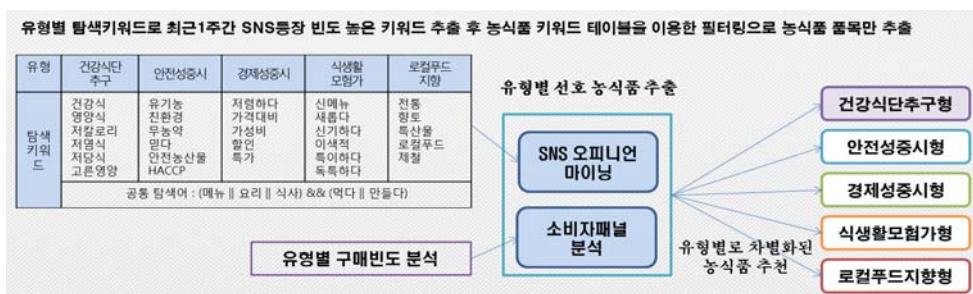
유형	건강식단 추구형	안전성 중시형	경제성 중시형	식생활 모험가형	로컬푸드 지향형
특징	<ul style="list-style-type: none"> <li>- 저칼로리 추구</li> <li>- 저염식 추구</li> <li>- 저당식 추구</li> </ul>	<ul style="list-style-type: none"> <li>- 유기농 선호</li> <li>- 원산지 확인</li> <li>- 인증(GAP, 무농약 등) 확인</li> </ul>	<ul style="list-style-type: none"> <li>- 저렴한 식재료 구입</li> <li>- 가격정보 확인</li> <li>- 식품선택시 가격고려</li> </ul>	<ul style="list-style-type: none"> <li>- 새로운 음식 경험선호</li> <li>- 이색적인 식당 선호</li> <li>- 새로운 요리 시도</li> </ul>	<ul style="list-style-type: none"> <li>- 전통음식 선호</li> <li>- 향토음식점 관심</li> <li>- 지역특산물 관심</li> </ul>

자료: 농림수산식품교육문화정보원, 2014

## <2> 5가지 유형을 기반으로 선호 농식품 추출

- 소비자패널 분석 결과와 SNS 오피니언 마이닝을 농식품 키워드 테이블을 이용한 필터링으로 농식품 품목만 추출
- 이를 통해 5가지 유형별 최근 한 주간의 인기 식품 결과를 얻을 수 있음

[그림] 소비자 유형별 농식품 추천 개념도 및 결과



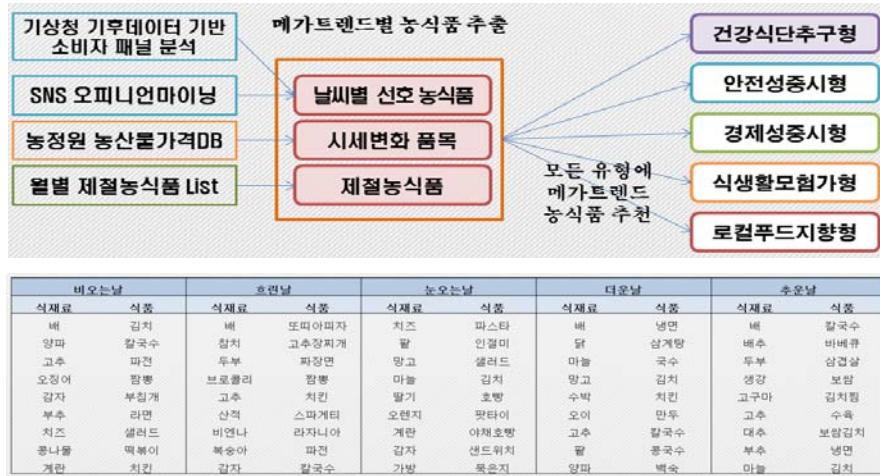
건강식단추구형		안전성증시형		경제성증시형		식생활모형가형		로컬푸드지향형	
식재료	식품	식재료	식품	식재료	식품	식재료	식품	식재료	식품
두부	샐러드	우유	샐러드	마늘	피자	치즈	샐러드	된장	김치
양파	김치	쌀	파스타	배	파스타	새우	크림파스타	돼지	비빔밥
우유	가슴살	버섯	시리얼	고구마	치킨	김치	치킨	갈비	국수
마늘	닭가슴살	사과	쌈밥	베이컨	튀김	마늘	스테이크	마늘	냉면
고추	멸치볶음	고추	불고기	양고	김치	배	떡볶이	나물	수육
계란	낫또	고구마	장아찌	양파	해장국	고구마	짬뽕	밥상	만두
버섯	어묵볶음	양파	청국장	고추	떡볶이	베이컨	감자튀김	식혜	국밥
된장	청국장	토마토	김부각	오경어	볶음밥	돼지	고르곤졸라파자	돼지고기	간장게장
닭가슴살	등갈비	치즈	두루치기	토마토	짬뽕	망고	칼국수	한우	육개장
소고기	된장찌개	당근	인절미	계란	토스트	소고기	바베큐	대추	깻갈

자료: 농림수산식품교육문화정보원, 2014

## <3> 메가트렌드별 농식품 추출

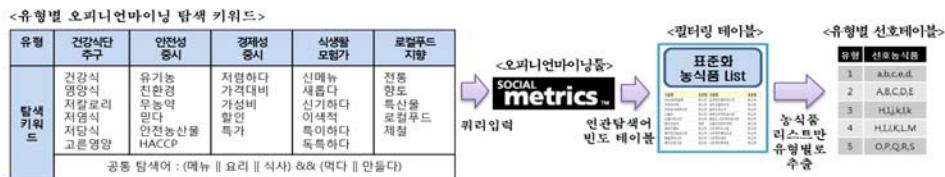
- 4년간('10-'13)의 날씨 데이터를 이용하여 특정 날씨와 해당 날짜 추출 후, 4년간 패널 구매 데이터에서 해당 날짜에 많이 구매한 농식품 추출
- 최근 3년간 도매시장 가격정보를 분석하여 급격한 시세변화 감지 알고리즘 개발 후 알고리즘에 의한 자동 추천
- 비오는날, 흐린날, 눈오는날, 더운날, 추운날의 5가지 메가트렌드에 따라 인기 있는 식품 리스트를 얻을 수 있음

[그림] 메가트렌드별 농식품 추천 개념도 및 결과



자료: 농림수산식품교육문화정보원, 2014

[그림] 소비자 유형별, 메가트렌드별 선호 농식품 추출을 위한 알고리즘



<오피니언 마이닝을 통한 추천 알고리즘(예시: 소비자 유형별 추출)>



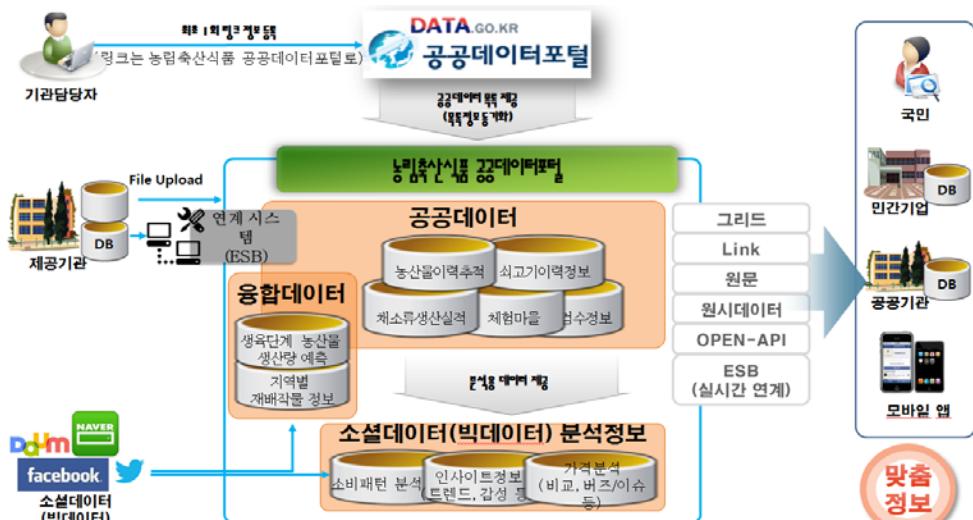
<소비자패널 데이터를 통한 추천 알고리즘(예시: 메가트렌드별 추출)>

자료: 다음소프트, 2014

■ 농림축산식품 공공데이터 통합 서비스 체계를 구축하기 위해 데이터 개방 및 포털 서비스 구축 등 데이터 공동 활용 체계 마련

- 농식품부 및 소속/산하 기관 등 15개 기관의 159종의 공공데이터와 연구 용역에서 발굴된 121종 데이터 등을 수집
- 수집된 데이터는 ESB솔루션<sup>11)</sup>을 통해 자동으로 DB화 되도록 구성하였고, 공공데이터 포털을 통해 GRID, Open API 등의 방식으로 서비스

[그림] 농림축산식품 공공데이터포털 목표 구성도



자료: 농림수산식품교육문화정보원, 2014

11) ESB 솔루션: Enterprise Service Bus Solution, 서비스들을 컴포넌트화된 논리적 집합으로 묶는 핵심 미들웨어이며, 비즈니스 프로세스 환경에 맞게 설계 및 전개할 수 있는 아키텍처 패턴

## 효과 및 향후 적용 확대 방안

- 서비스 정교화 실현 및 모바일 수요 대응 등 지속적인 서비스 개선 노력을 통한 이용 확대
  - 농식품 공공데이터의 품질관리, 가격 데이터 보정, 포털 및 빅데이터 서비스 기능 보완, 연계 솔루션 테스트 진행 등 지속적인 서비스 정교화를 통해 시시각각 변화하고 추가되는 정보에 대응할 계획
  - 사용자의 활용성을 높이기 위해 향후 모바일 어플리케이션을 통한 서비스 제공 예정
- 공공데이터를 활용한 다양한 서비스 모델 개발의 대표 사례로 향후 다양한 기관의 데이터 추가 개방 및 이에 대한 활용 활성화

## 20. 조류 인플루엔자(AI) 확산 조기대응

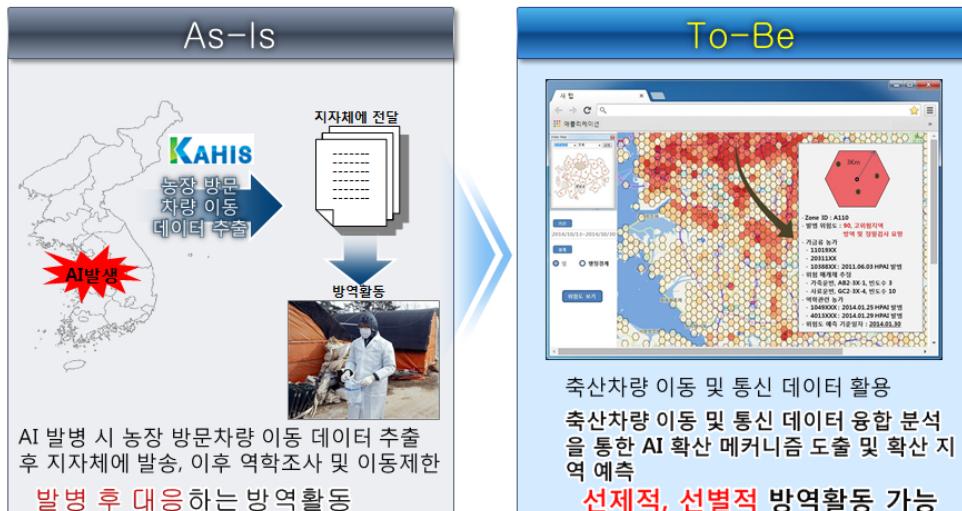


국가동물방역통합시스템(KAHIS) 데이터와 KT의 통화로그 데이터를 연계 분석하여 조류인플루엔자(AI)의 확산 경로 예측

### 추진 목적 및 배경

#### ■ 사업 추진의 배경

- 현재 팀문조사와 방역활동 등 AI 발병 후 대응을 중심으로, 광범위 확산에 대한 효과적 방지에 한계
- KT 기지국 통계 및 농림축산검역본부 KAHIS 축산차량 이동 데이터를 활용하여 AI 확산 지역 예측 모델을 개발



## ■ 사업 추진의 필요성

- AI학산의 명확한 규명을 통한 2차 감염피해 예방
- 축산농가의 막대한 피해액 최소화
- 효율적 방역을 통한 국가 예산의 절감
- 빅데이터 기반 농축산 분야 ICT 융복합 촉진을 통한 공공 이익 극대화

## 추진 내용

| 참여기관 | (주)KT, 농림축산검역본부

## | 활용데이터 |

구분	데이터	데이터 양	제공 기관
통화로그 (CDR)	기지국 정보	67TB/월	KT
	기지국별 통화량 통계 정보		
KAHIS-축 산차량, 농장정보	축산차량 농장방문 기록 정보	200MB/월	농림축산검역본부
	전국 농장 정보	90MB/월	
	AI 발병 농장 정보	70MB/주	
	축산 차량 등록 정보	560MB/주	

## | 분석 내용 및 기법 |

### 데이터 처리 및 분석 기법

## ■ DATA 수집 · 전처리

- KT 데이터 분석 전문가가 내부 분석 플랫폼 활용하여 데이터 전처리, 통계 테이블 생성 등 분석을 위한 형태로 데이터 가공 작업 진행

### ■ 가금류 축산업 관련자 인터뷰 수행

- 관련자 심층 인터뷰를 통해, AI 확산 매커니즘 규명 및 예측모델 개발에 활용
- 인터뷰 대상: 동물 감염병 전문가, 가금류 축산업 종사자, 가금류 농장주, 가축 운반, 사료운반 등 기능별 관련 종사자, 방역 담당 공무원 등

### ■ AI 확산 매커니즘 규명

- AI 발생 농가별 발생원인 및 확산 매개체 분석
- 기지국 통계 데이터와 가금류 농장 관련자들의 차량이동데이터를 분석하여 농가별 발생원인 분석
- AI 확산의 핵심 요인으로 추정되는 확산 매개체 선정
- AI 확산 관련 핵심 요인 추출
- 발병 농가 사이의 연관성 분석

### ■ GIS 기반 분석 및 시각화

- KT의 GIS 전문가 인력 투입하여 확산 매커니즘 분석 결과를 지도상에 시각화
- 공간분석, 방문 매개체 분석, 거리측정, 농가 위험도 분석 및 시각화 작업 수행

### ■ 데이터 분석 과정

#### ① AI 확산 예측모델 개발

- KT 기지국 통계 데이터 및 농림축산검역본부 KAHIS 축산차량 데이터를 통한 차량·사람 이동과의 연관성 분석
- 발병 농가 특성 및 발병 농가 방문 차량 특성별 연관성 분석 확산 매개체의 핵심 요인 선별 및 가중치 부여
- 이를 적용한 AI 확산 예측 모델 개발

## ② AI 확산 예측모델 정확도 검증

- 과거 AI 발병 이력 데이터를 활용한 가상 예측 시뮬레이션 수행
- 예측 모델에 의한 결과를 실제 발병이력과 비교하여 성능 검증

## ③ AI 확산 위험지역 선별 제공

- 농림축산검역본부 위기대응센터에 일별 확산 예상지역 지도 및 방역 대상 농가 리스트 전달
- Web 기반 GIS 시작화 및 레포트 제공

[그림] AI 확산경로 예측 및 대응 프로세스

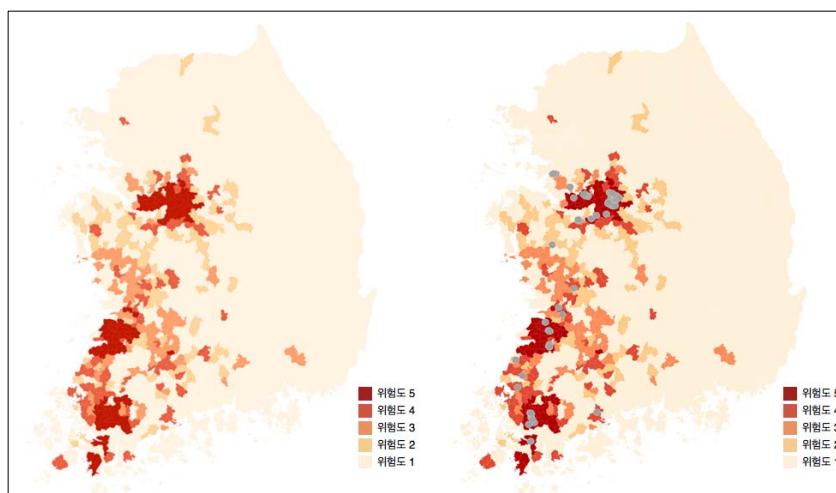


## 주요 분석 결과 및 활용방안

### | 주요 분석 결과 |

- 가축전염병의 주요 전염확산 요인은 차량으로 차량 이동경로를 따라 질병이 확산되는 것을 확인
  - 사료운반, 가축운반, 분뇨차량, 시료채취·방역 차량 등 차량 유형별 확산 확률이 다름
  - 이를 차량의 이동 데이터를 토대로 지역별 전염 위험도 시뮬레이션
- AI 발병농가의 약 83%가 고위험 추정지역(위험도5 : 전국영토의 약 4%) 내에서 발생한 것을 확인

[그림] 발생지 예측결과와 실제 발생지역

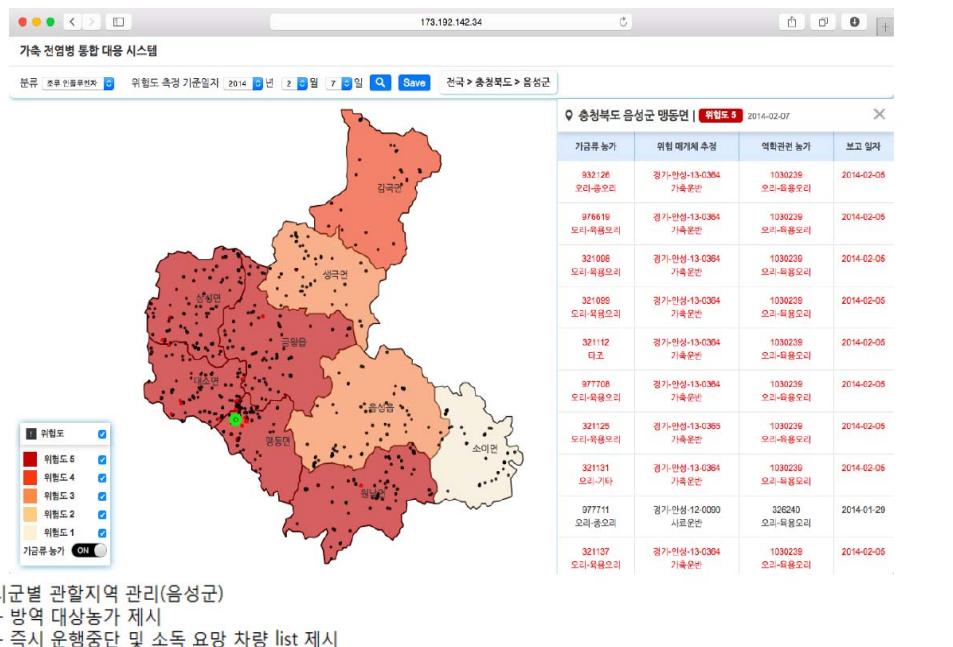


## | 서비스 계획 및 활용방안 |

### ■ 매월·필요시 분석리포트를 농림축산검역본부에 제공

- 향후 가축전염병 확산예측 시스템으로 확대 구축하여 운영하여 지자체 공무원이 방역을 위한 세부 행동계획을 제시

[그림] 가축 전염병 확산예측 시스템



시군별 관할지역 관리(음성군)  
- 방역 대상농가 제시  
- 즉시 운행중단 및 소독 요망 차량 list 제시

## 효과 및 향후 적용 확대 방안

### ■ 가축전염병에 대한 초동대처가 가능해져 확산 조기방지 및 피해최소화

※ 구제역, 돼지콜레라 등 타 전염병으로 확대 가능, 일본·중국 등에 서비스 모델 수출 가능

### ■ 타 사람/가축 전염병 확산에 대한 확대 적용

- 현재 발병/확산되고 있는 구제역과 같은 피해 규모가 큰 가축 전염성 질병에 확대 적용
- 가축 전염병은 AI 적용사례와 유사하므로 축산 차량 이동분석을 통해 확산 지역 예측가능

### ■ AI 확산 예측모델 자동화를 통한 실시간 대응체계 구현

- 서버 내 AI 확산 지역 예측 모델 적용 모듈 개발
- 확산 위험지역을 확인할 수 있도록 지도상의 표현기능 제공(Web 기반)
- 확산 위험 지역 리스트 확인 가능 리포트 daily 제공

## 21. 국도 비탈면 붕괴사고 예측



급경사지 지형 데이터와 기상정보를 연계 분석하여 국도 비탈면의 위험도 산정 모델 개발 및 예측 서비스 제공

### 추진 목적 및 배경

#### ■ 사업 추진의 배경

- 우리나라는 매년 장마철이나 태풍 내습과 관련된 집중호우에 의해 산사태나 비탈면 붕괴가 빈번하며, 최근 15년간 매년 평균 53번의 비탈면 붕괴로 인해 20여명의 사상자가 발생함



## ■ 사업 추진의 필요성

- 강우는 급경사지 붕괴의 중요요소임에 따라, 실제 강우가 많이 발생하는 우기철에 급경사지의 불안정화로 인한 붕괴가 빈번히 발생하고 있음
- 강우침투에 의한 급경사지 안정해석 모델 연구가 많이 진행되어 있으므로 이러한 모델을 기반으로 기상자료와 연계한 급경사지 특성, 사고이력에 대한 빅데이터 분석을 통해 사고예측 필요

## 추진 내용

|참여기관| 대한지적공사, 한국건설기술연구원, 모바일팩토리, SKT

### |활용데이터|

구분	데이터	데이터규모	보유기관
도로비탈 면데이터	도로비탈면 기초조사자료	6,587 지점	건설기술연구원
	도로비탈면 정밀조사자료	936 지점	
강우량등 기상정보	일강수량, 3일 누적 강수량, 7일 누적 강수량	1년간 데이터	기상청
	일평균기온	1년간 데이터	

### |분석 내용 및 기법|

#### ■ 데이터 처리 및 분석 기법

- 도로비탈면유지관리시스템의 기초 및 정밀 조사 데이터 분석
- 기상조건 및 공간정보 인자 도출
- 빅데이터 분석 기법을 활용한 비탈면 붕괴 위험 지수 모델 개발

- 위험 경보 서비스를 위한 위험 등급 설정 및 지오펜싱(geofencing)<sup>12)</sup> 기술 개발
- 실시간 기상 정보 연계 비탈면 위험 예측 경보 모바일 서비스 개발

[그림] 비탈면 사고 예측 시스템 구성도

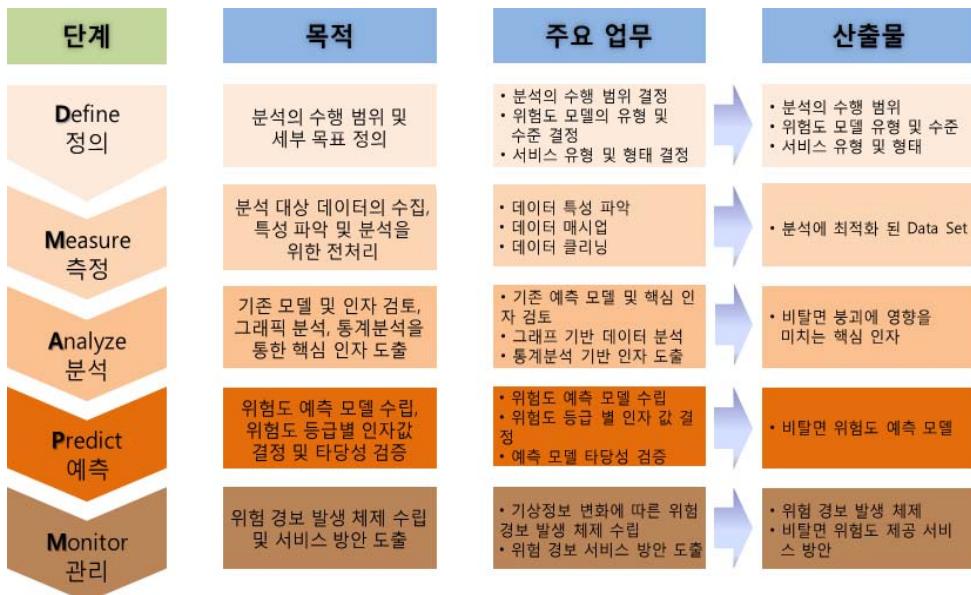


## ■ 데이터 분석 과정

빅데이터를 활용한 국도 비탈면 봉괴 위험도 예측 모델 수립 및 서비스 방안 도출은 DMAPM 5단계를 거쳐 수행됨. DMAPM 단계는 분석의 수행 범위 및 세부 목표를 정의하는 Define 단계, 분석 대상 데이터의 수집 및 특성 파악과 전처리를 수행하는 Measure 단계, 기존 예측 모델 및 핵심 인자 검토 그리고 새로운 핵심 인자 도출을 위한 그래픽분석 및 통계분석을 수행하는 Analyze 단계, 분석된 결과를 토대로 위험도 예측 모델 수립 및 위험도 등급별 인자 값 결정과 타당성 검증을 실시하는 Predict 단계, 봉괴 위험도 예측 모델을 활용한 기상정보 변화에 따른 위험 경보 발생 체제 수립 및 서비스 방안 도출을 위한 Monitor 단계로 구성

12) Geographic과 Fencing의 합성어로 특정 구역에 대한 사용자 출입현황을 알려주는 API

[그림] 데이터 분석 과정



### ① Define 단계: 분석의 수행 범위 및 세부 목표 정의

- 국도 비탈면을 여러 특성(지질학적 특성, 환경적 특성, 지리적 특성 등)에 따라 구분하고 이를 바탕으로 실제적인 분석의 수행 범위를 결정함. 또한 기존의 비탈면 봉괴 위험성을 평가할 수 있는 비탈면 재해 예측법 및 기준을 고찰하여 이러한 방법론의 한계점을 파악하고 개선방안을 도출
- 최종적으로 분석하여 개발할 비탈면 위험도 모델의 유형(수리적 모형, 인과형 모형 등)을 결정하며 개발되는 모델의 수준(모델의 정확도, 정밀도 등)을 결정하여, 이를 토대로 제공되는 서비스의 유형(위험도 수준에 따른 경보 서비스, 위험도 확률 제공 서비스, 실시간 위험도 제공 서비스 등)과 제공되는 형태(PUSH알림, 봉괴 위험 지도 등)를 결정

## ② Measure 단계: 도로비탈면유지관리시스템 데이터와 기상 데이터의 매쉬업

- 도로비탈면유지관리시스템의 비탈면 정밀조사 데이터와 기상청에서 제공하는 국가기후자료센터로부터 획득한 기상 데이터를 위도와 경도를 기준으로 매쉬업 실시
- 데이터 매쉬업을 통해 국도 비탈면 붕괴 위험도 분석에 비탈면의 고유 특성을 나타내는 인자(비탈면 관리시스템의 정밀조사 데이터)와 실시간으로 변화되는 인자(기상청의 기상 데이터)를 동시 고려하는 실시간 붕괴 위험도 예측 모델 수립
- 분석의 용이성을 위한 데이터 정제(Cleaning) 및 전처리(Preprocessing)

## ③ Analyze 단계: 그래프 분석 및 문헌조사를 통한 붕괴 주요 인자 도출

- 데이터를 붕괴 레코드와 미붕괴 레코드로 분류하여 각각 그래픽 분석을 실시. 분석되는 인자의 데이터 종류가 연속형인 경우 Box Plot을 활용하고 범주형의 경우 Dot Plot을 사용하여 붕괴와 미붕괴 간에 확연한 차이 보이는 인자들을 선별한 후 통계분석으로 확정
- 비탈면 붕괴와 관련된 50편의 논문에 대한 문헌조사를 통해 비탈면 붕괴에 영향을 미치는 인자들을 확인하고 이를 토대로 비탈면 붕괴에 유의한 후보 인자 리스트를 작성함. 또한 비탈면 붕괴를 예측하는 기존의 예측 모델을 검토

## ④ Predict 단계: 데이터 분석 기법

- 붕괴 이력 기반 빅데이터를 활용한 비탈면 분석 모델 도출
- 연관분석을 통한 붕괴 패턴 파악, 군집분석 및 의사결정나무 분석을 통한 패턴 도출과 전체 레코드의 군집화
- 의사결정나무 분석을 이용한 군집별 붕괴와 미붕괴 분류분석
- 기상정보에 따른 비탈면 위험도 등급 제시

- 최종 위험도 산정 모델에 대해 기존 data set 및 외부 data set을 활용하여 모델 검증을 실시하고 정확도를 개선할 수 있는 방안 제시

#### ⑤ Monitor 단계: 서비스

- 비탈면 위험도 경보 서비스를 위한 분석 결과 DB화
- 위험 경보 발생 순서도 작성 및 시스템 흐름도 제시

### 주요 분석 결과 및 활용방안

#### |주요 분석 결과|

##### ■ 봉괴여부에 영향을 미치는 인자 도출

- 강우량, 비탈면의 수분 함유량, 지하수 맥 여부, 풍화도 등이 봉괴에 영향을 미치는 주요 인자로 판명

##### ■ 인자들 간 영향도 분석(상관분석)을 수행하여 봉괴/미봉괴의 특성 패턴화

[그림] 봉괴영향인자 분석결과



## | 서비스 계획 및 활용방안 |

### ■ 대한지적공사의 대국민 모바일 서비스 앱 “LX토지알림e”에 비탈면 위험알람서비스 및 주변 비탈면 정보 조회서비스를 적용하여 배포

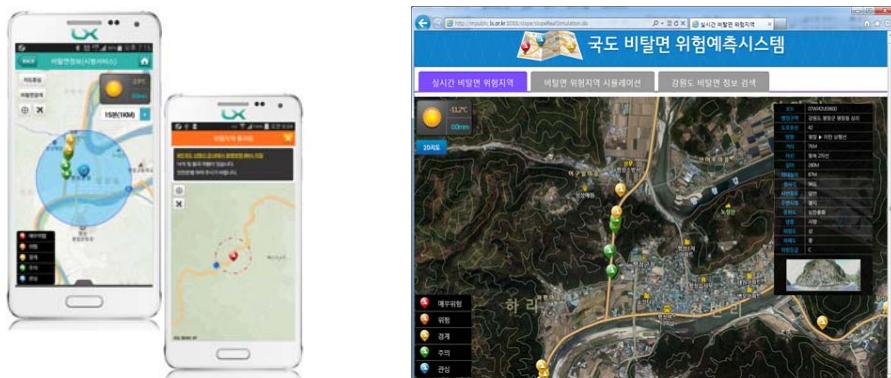
※ 구글 Play 스토어 및 앱 스토어(iOS)를 통해 배포(2015년 2월)

※ 인터넷기사, 홈페이지, SNS, 블로그 등 다양한 매체를 통해 홍보 추진

- 향후 T맵 등 내비게이션 서비스와 협력하여 우회경로 안내 서비스 추진
- 한국건설기술연구원과 대한지적공사가 협업체계 구축하여 국토교통부 도로 운영과에 정책 제안
- 강원영서 시범사업 지역 외 지자체로 확대 추진
- 지자체 2016년 예산 반영 추진

### ■ 활용 서비스

[그림] 국토비탈면 붕괴사고 예측 서비스



#### 대국민 모바일 서비스

: 주변 비탈면 위험정보 조회 및 지오펜싱 기술을 활용한 위험지역 진입 시 위험알람 서비스 제공

#### 국도 비탈면 위험예측시스템

: 기상정보와 연계하여 실시간 비탈면 붕괴 위험도 조회, 시뮬레이션, 정보 검색으로 관리체계 고도화 가능

- 예측 모델 검증 후 비탈면 보강공사 우선순위 결정 등 의사결정에 활용
  - 3차원 지형분석기술 활용 등 붕괴예측 서비스 고도화

## 효과 및 향후 적용 확대 방안

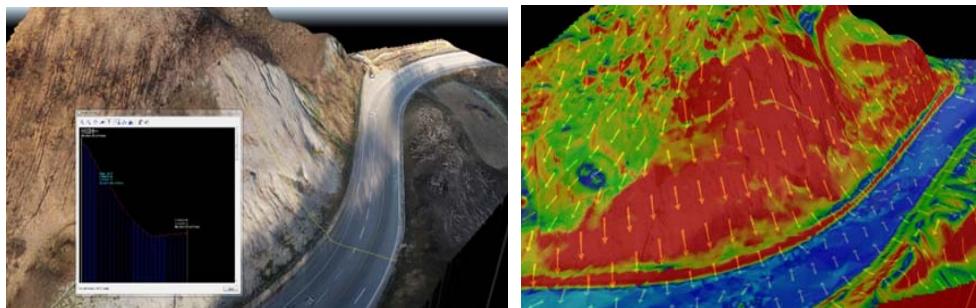
- 성과 및 시사점
  - 공공데이터와 민간데이터가 융합되어 가치있는 정보를 국민에게 제공하여 안전사회 구현
  - 과학적인 비탈면 보강공사 우선순위 선정에 활용하여 예산 절감
  - 지자체가 관리하는 지방도에 대한 실태조사 추진과 예측 모델 적용으로 비탈면 관리사업 확산
  - 국가재난재해서비스와 연계한 서비스 확대(추후 관계 기관 협의를 통해 반영)
- 향후 발전방향
  - 국토비탈면 붕괴관련 모바일 활용 시범서비스의 제공 후 분석 데이터와 서비스 종류를 확대하고 서비스 확산 및 홍보

[그림] 국도비탈면 관리 서비스의 발전방향



● 3차원 지형분석기술을 활용하여 고도화된 붕괴예측 서비스 추진

[그림] 국도비탈면 붕괴예측서비스(안)



### 참고문헌

BoA, 수익성 및 업무효율 제고를 위한 빅데이터

SAS, *'Big Data for the Next Big Idea in Financial Services'*

SAS(2013. 5), *'Big Data in Big Companies'*

SAS, *A comprehensive environment for more efficient forecasting*

Information Management(2013. 4. 30), *Bank of America CIO on Big Data, Emerging Enterprise Tech*

Forbes(2013. 6. 11), *Banks Use Big Data To Understand Customers Across Channels*

허츠, 실시간 VOC 분석으로 고객 만족도 향상

Teradata(2014. 1. 27), *Hertz: Finding Gold in Integrated Data*

IBM, *How big data is giving Hertz a big advantage* ↴

Customerintelligence360(2014. 1. 9), *Secret Big-Data Marketing Helps Hertz Drive Money-Spinning Upgrades*

GS홈쇼핑, 고객 추천 서비스 정교화

그루터(2013), *Use case summary*

ZDNETkorea(2013. 3. 19) GS홈쇼핑, 빅데이터 제대로 써보니…

디지털데일리(2013. 5. 8), [창간8주년/대한민국 빅데이터] GS홈쇼핑 '빅데이터 기술 새재화 주력'

디지털타임스(2013. 3. 19), GS샵 빅데이터 선도적 도입 주목

롯데백화점, 고객 세분화를 통한 타깃 마케팅

롯데백화점(2014. 3), 「롯데백화점 Big Data를 만나다」

매일경제(2014. 6. 2), *롯데백화점, 고객 쇼핑패턴 파악…연령별 맞춤 마케팅*

Ancestry.com, 적극적 데이터 수집을 통해 온라인 가계도 서비스 구현

*MapR, case study 'Ancestry.com'*

*Ancestry.com, 'Managing Big Data Reaching Back to the 11th Century'*

*Ancestry.com Blog, Adventures in Big Data: How AncestryDNA Uses Hadoop and HBase*

*Fiercebigdata(2014. 8. 4), How Ancestry.com uses big data*

*Bill Yetman, <http://blogs.ancestry.com/techroots/>*

오비츠, 사용자 특성을 파악하여 맞춤 검색 결과 제공

*Orbitz, Architecting for Big Data- Integrating Hadoop into an Enterprise Data Infrastructure*

*tnoz(2014. 1. 28), Orbitz Labs debuts with experimental Big Data travel tools for consumers*

*LifeHacker(2013. 1. 7), How Web Sites Vary Prices Based on Your Information*

*International, Business Times(2014. 11. 12), Mac and Android Users Charged More on Shopping Sites Than iPhone and Windows Users*

NC소프트, 게임 내 사기 탐지 시스템 구현

*NC소프트(2014), 'Data Analysis for Game Fraud Detection'*

멜론, 이용자 관심도에 따른 콘텐츠 추천

*Loen Entertainment(2014), MelOn 빅데이터 도입사례*

*스포츠동아(2014. 6. 24), 멜론, 빅데이터 이용 아티스트 이용자 직접 연결, 세뮤직 콘텐츠 플랫폼 진화*

*ZDnet korea(2014. 7. 17), 멜론이 빅데이터 기술로 기대하는 3가지*

*서울경제(2014. 7. 7), [서울경제 TV SEN] '빅데이터'의 힘..맞춤형 콘텐츠 인기*

*아이뉴스24(2014. 7. 16), 멜론의 유쾌한 도전 '맞춤형 선곡도 빅데이터로'*

## UNC헬스케어, 환자의 재입원 비용 절감

*IBM Information Ondemand (2011), IBM Medical Records Text Analytics Solution Helps UNC Healthcare Improve the Quality of Hospital Discharges*

*IBM, 'how big data analytics reduced Medicaid re-admissions.'*

*IBM, 'UNC Health Care System: Bringing Health Informatics to a New Level.'*

*IBM Data Summit (2014), Big Data and Healthcare: Key Technologies and Strategies*

*IBM(2013. 10. 11), UNC Health Care Uses IBM Analytics to Manage Medical Data and Improve Patient Care*

*EnterpriseTech(2013. 10. 11), UNC Health Care Selects IBM Analytics for Patient Care*

*BusinessCloudNews(2013.11.13.), How UNC Health Care built a big data platform that saves lives*

## 서울아산병원, 의료연구 편의성 확대

*Microsoft(2014), 서울아산병원 국내 최초 '개인정보 보호법과 '생명윤리 및 안전에 관한 법률'을 준수하는 연구 정보 검색 시스템' 가동*

*메디컬옴저버(2014. 11. 6), 서울아산병원은 왜 '빅데이터'에 집중할까?*

*디지털타임스(2014. 10. 23), 한국MS, 서울아산병원에 빅데이터용 DW 공급*

*이데일리(2014. 10. 23), 서울아산병원, 마이크로소프트 차세대 DW 도입!*

## GE, '지능형 항공 운영' 서비스

*GE, 'Flight Efficiency Services'*

*GE Aviation(2013. 10), 보도자료 'GE Aviation Announces Customers for Flight Efficiency Services'*

*GE(2013), 'Fuelling Global Airlines - GE's Industrial Internet makes aviation sector more efficient.'*

*GE Software(2013), 'Big Data and the Industrial Internet'*

*GE(2012. 11. 26), 'Industrial Internet: Pushing the Boundaries of Minds and Machines'*

## 볼보, 운행 정보 활용한 자동차 안전 실현

CIO 매거진(2014. 10. 27), **볼보 빅데이터 사례, 전사적 데이터 활용에 초점을 맞춘다**

IBM Big Data & Analytics Hub(2013), *Big Data Pioneers: Volvo Case Study*

Volvo news(2013. 12), *Big data can improve business*

Teradata(2012. 5), *『A car company powered by data』*

## 캐터필러, 직원 및 기기 데이터 분석을 통한 제조 생산성 향상

IBM software case study, *『Connecting employee engagement and key metrics impacts the bottom line for Caterpillar』*

Reuters(2014. 3. 20), *From dumb iron to Big Data: Caterpillar's dealer sales push*

Kenexa.com(2013. 1. 8), *how-caterpillar-tied-employee-engagement-to-business-outcome*

IBM, case study for Caterpillar, *Employee Engagement and Organizational Performance: What Caterpillar Discovered*

## 한국남동발전, 발전설비 운영효율 극대화

한국남동발전(2014. 4), *『빅데이터를 활용한 발전설비 운영효율 극대화 방안』*

매일경제(2014. 6. 2), *한국남동발전, 자원 최적 관리로 연료비 절감 성과*

디지털데일리(2013. 4. 25), *한국남동발전, 빅데이터 분석 도입해 경영혁신 나서*

## 농림수산식품교육문화정보원, 스마트 농정 실현을 위한 빅데이터 서비스 기반 구축

농림수산식품교육문화정보원(2014), *『농림축산식품 공공데이터 포털 및 빅데이터 서비스』*

농림수산식품교육문화정보원(2014), *농림축산식품 공공데이터 포털 및 빅데이터 서비스 기반 구축 사용자 매뉴얼*

\* 본문의 사례는 해외, 국내, 2014년 미래창조과학부 시범사업으로 구성되었음을 알립니다.

## 2015년 빅데이터 글로벌 사례집

2015년 5월 인쇄

2015년 5월 발행

| 발행인 : 서 병 조

| 발행처 : 한국정보화진흥원 미래전략센터

| 집 필 : 신신애, 김성현, 송경빈

서울시 중구 청계천로 14

TEL : (02)2131-0114

| 인 쇄 : 전우용사총(주) TEL : (02)426-4415

<비매품>

