2020 IT 21
Global Conference
Digital New Deal
Technology Essentials
디지털 뉴딜 기술 핵심

Session 6-2

셀프 서비스 BI를 위한 빅데이터 분석 플랫폼

하상윤 실장 (티맥스)

[요약문]

모두가 AI와 빅데이터를 관심을 가지고 있고 이를 활용해 비즈니스 인사이트를 도출하고자 하지만 전문가 부족, 데이터 가공의 어려움 등으로 실제 도입에 어려움을 겪고 있음. 이를 극복하기 위해 비전문가도 쉽게 데이터를 분석할 수 있는 분석 자동화와 자동화된 기계학습 개념을 소개하고 데이터를 활용할 수 있는 방법들을 알아보고자 함

목차

1. 추진 배경 및 목적: 셀프 서비스 패러다임, 데이터 분석 성공 요건
2. 빅데이터 분석 현안 및 이슈: 기계학습의 어려움, 데이터 사이언티스트 부족 등
3. 빅데이터 분석 방향성: 분석 자동화의 필요성과 방향성
4. 빅데이터 분석 자동화 구현방안을 위한 핵심 키워드: 데이터 수집체계부터 분석에 대한 일관된 파이프라인, 현업이 고민하고 개선할 수 있는 손쉬운 분석 환경 제공

[발표자 약력]

2005년	고려대학교 컴퓨터학 학사
2005년~2013년	삼성SDS 정보통신기술 연구원
2014년	코리아폴리스쿨 스마트개발팀 팀장
2015년~현재	티맥스 기술 실장

관심분야 : 빅데이터 Analytics, 클라우드 컴퓨팅, 웹 개발, Framework, Interface 등

Better Cloud, Better Tomorrow

셀프서비스 BI를 위한 빅데이터 분석 자동화와 데이터 활용

2020. 09.

Tmax

Contents

1. 추진 배경 및 목적 – Self Service Paradigm

"빅데이터, 인공지능의 급부상으로 현업 사용자가 데이터를 통합/분석하려는 니즈가 점차 증가 "

"DO IT YOUR SELF, YOUR WAY"

End User Computing

Self Service BI

1. 추진 배경 및 목적 – Self Service BI

"의사결정을 가속화 하는 Self Service BI "

출처 : 팍사타

1. 추진 배경 및 목적 – 데이터 분석의 성공 요건

데이터 전문가와 업무 전문가가 협업할 때
가장 유의미한 시너지를 낼 수 있지만 매우 어려운 과제

수많은 숫자 앞에서
어떤 일을 해야 하는지 모르는

현업

비전문가가 손쉽게 분석이 가능한 환경

현장의 목소리를 알지 못하는

데이터 사이언티스트

데이터의 추출-전처리 과정을 획기적으로 감소할 수 있는 방안

데이터 기반 의사결정을 가속화하는 방향성이 필요

2. 빅데이터 분석 현안 및 이슈

➢ 분석자동화 이슈 및 시사점

- 국내 기업의 빅데이터 이용률은 9.5%, 한국의 빅데이터 활용률은 63개국 중 31위 차지(한국정보화진흥원/IMD, '18)
- 빅데이터 미도입 이유로는 전문인력부재 41.5%, 데이터 부재 33.7%, 작은 기업 규모 26.9%가 가장 많이 꼽힘(한국데이터산업진흥원, '18)
- 빅데이터 활용을 진흥하기 위해서는 기술력뿐만 아니라, 전문가에 의존하지 않는 IT 환경이 마련되어야 함.

빅데이터 분석 자동화의 어려움	빅데이터 기술 지향점
• 원천 데이터 부재 • 데이터 사이언티스트와 같은 전문 인력 부족 • 개개인의 전문가에 의존하는 환경으로 인한 기업 고유의 데이터 분석 노하우 미비	• 현업자도 적시에 수행할 수 있는 데이터 분석 기술 • 탐색적 데이터 분석 및 분석 모델링 자동화 기능으로 손쉬운 분석 결과 제공

3. 빅데이터 분석 방향성

➢ 분석 자동화 필요성 및 도입 방향성

데이터 분석 수요 증가

데이터 분석 수요는 증가하고 있으나, 전문가 인력이 부족한 현상과 데이터 분석 복잡도는 계속해서 상승하고 있음

➡ 데이터 분석 수요 증가에 대비한 기술은 **빅데이터 특징 인지 필요**

분석 모델 생성 어려움

기업/기관 내 컴퓨팅 환경은 안정적인 환경이지만, 데이터 분석 모델 탐색이나 모델 추천이 쉽지 않음

➡ 적합한 분석 모델을 찾기 위해 **효율적 분석 모델 탐색 및 추천 알고리즘 필요**

손쉬운 분석 문제 해결

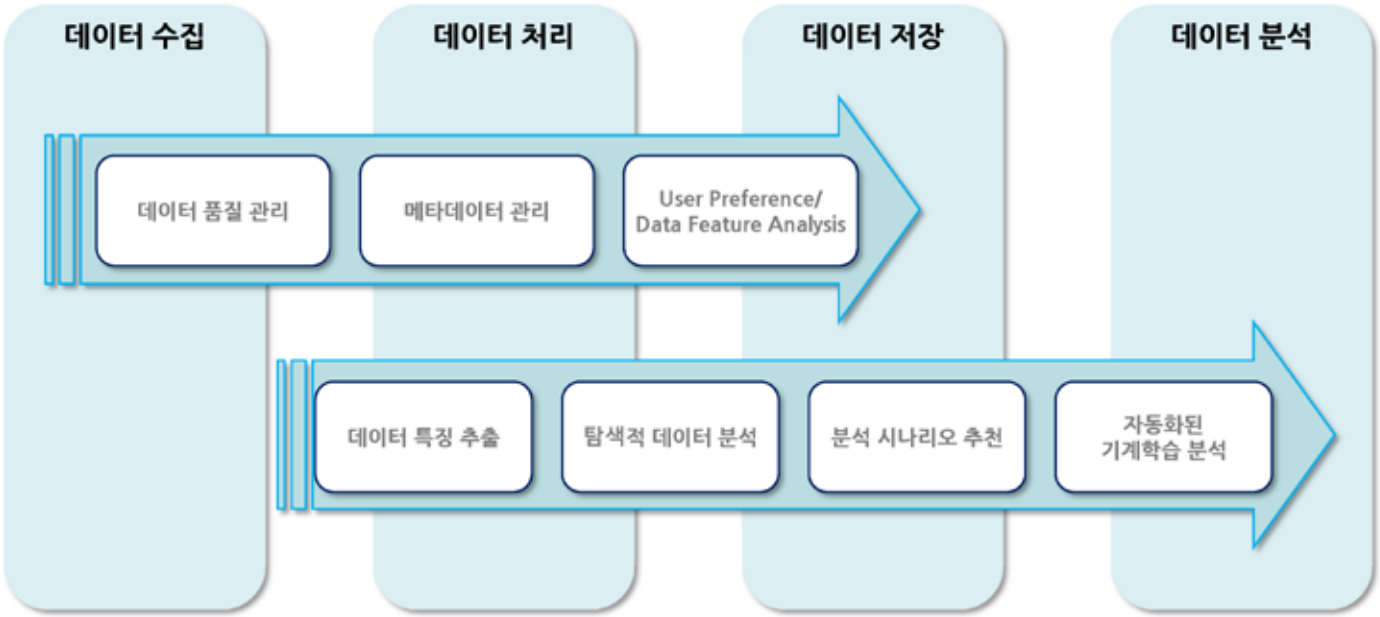
인공 지능 기술을 다룰 수 있는 인력의 부재로 인공지능을 활용하기 쉽지 않음

➡ 누구나 쉽게 데이터 분석을 하기 위해 **분석 모델 추천 자동화 필요**

"비전문가도 쉽게 분석할 수 있는 End User Computing"

4. 빅데이터 분석 자동화 구현 방안을 위한 핵심 키워드

"분석에 소요되는 복잡한 과정을 자동화/지능화하여
일반 사용자도 손쉽게 고급 분석이 가능한 환경"



4. 빅데이터 분석 자동화 구현 방안 > ① 데이터 품질 관리

➢ **빅데이터 환경과 같은 서로 다른 주체들은 관리 체계가 다양하여 품질 저하의 우려가 있고 모델의 성능에 악영향을 미치기 때문에 체계적인 데이터 품질 관리가 필요**

품질관리 대상 선정의 중요성

➢ **다수 사용자/부서/시스템에 동일기준으로 적용되어야 하는 데이터**

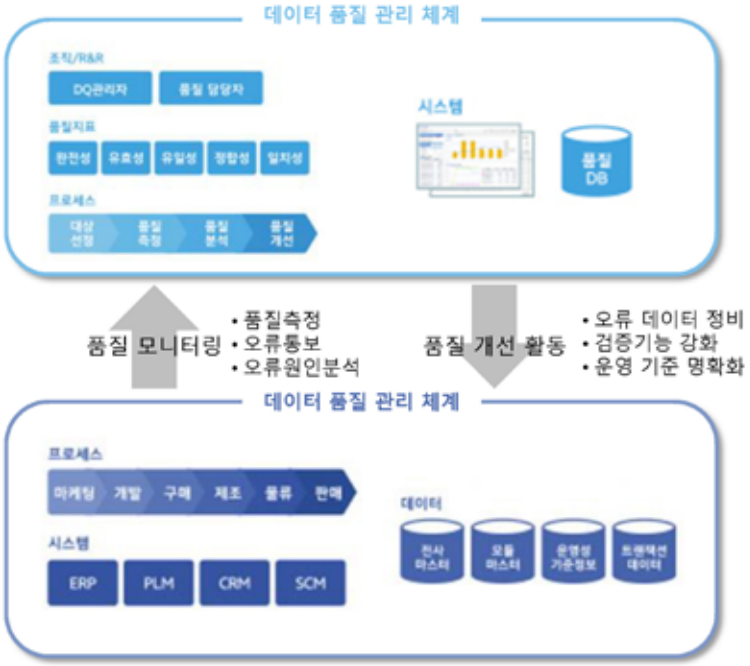
- 품질 이슈로 인한 영향이 크므로 데이터 품질개선 활동에 대한 효과도 매우 큼

➢ **자주 변동되지 않고 장시간에 걸쳐서 활용되는 데이터**

- 업무 과정에서 빈번하게 등록, 변경되는 데이터는 대부분 트랜잭션 데이터
- 품질관리는 비즈니스 프로세스 개선 및 시스템 보완 영역에 가까우며, 기준정보의 경우 정적인 데이터로 오랜 시간 동일값을 유지하는 경우가 많음

➢ **품질 수준을 정량화된 형태로 측정할 수 있는 데이터**

- 정보의 누락, 실물과 정보의 불일치, 시스템간 불일치와 같은 유형으로 명확하게 오류여부 판단이 가능해야 함



4. 빅데이터 분석 자동화 구현 방안 > ① 데이터 품질 관리

➢ 데이터 품질 측정은 빅데이터 품질의 선행 작업으로 이를 구조화, 계량화하여 빅데이터 분석 결과의 정확도를 높여야 함

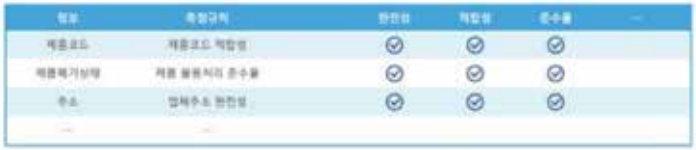
주요 품질 지표

품질지표	데이터 오류	설명
완전성	정보의 누락	필수 속성은 반드시 데이터 값이 채워져 있어야 함 - 예) 고객명, 모델 Spec 정보 필수
유효성	형식, 산식 등의 규칙 오류	데이터 값이 업무규칙을 준수하여 업무적으로 의미 있는 값이어야 함 - 예) 코드값은 사전 정의된 목록 내에 존재해야 함
유일성	동일 데이터 중복	동일 데이터는 중복없이 하나로 관리되어야 함. - 예) 모델 1 Spec 1 Code
정합성	연관정보의 일관성 오류	상호 관련이 있는 테이블, 속성 간 데이터 값이 모순되지 않아야 함. - 예) 모델 1 Spec 1 Code
일치성	시스템간 불일치	정보 수신시스템은 연계받은 정보를 임의적으로 삭제, 수정없이 활용해야 함. - 예) MDM과 ERP간 동일제품코드에 대한 정보 일치성
적시성	정보 연계 지연	정보 활용시스템에 업무적으로 적시에 제공되어 활용 가능하여야 함. - 예) MP 수행 이전 시점에 확정 Demand 연계

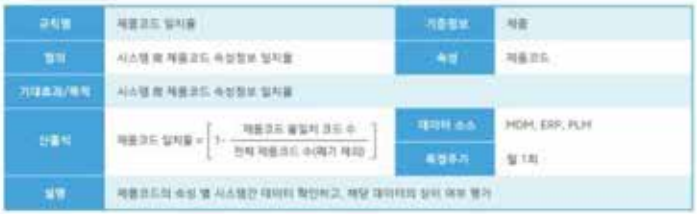
Step1. 데이터 품질 지표 정의

완전성 | 유효성 | 유일성 | 정합성 | 일치성 | ...

Step2. 정보별 적용 지표 선정

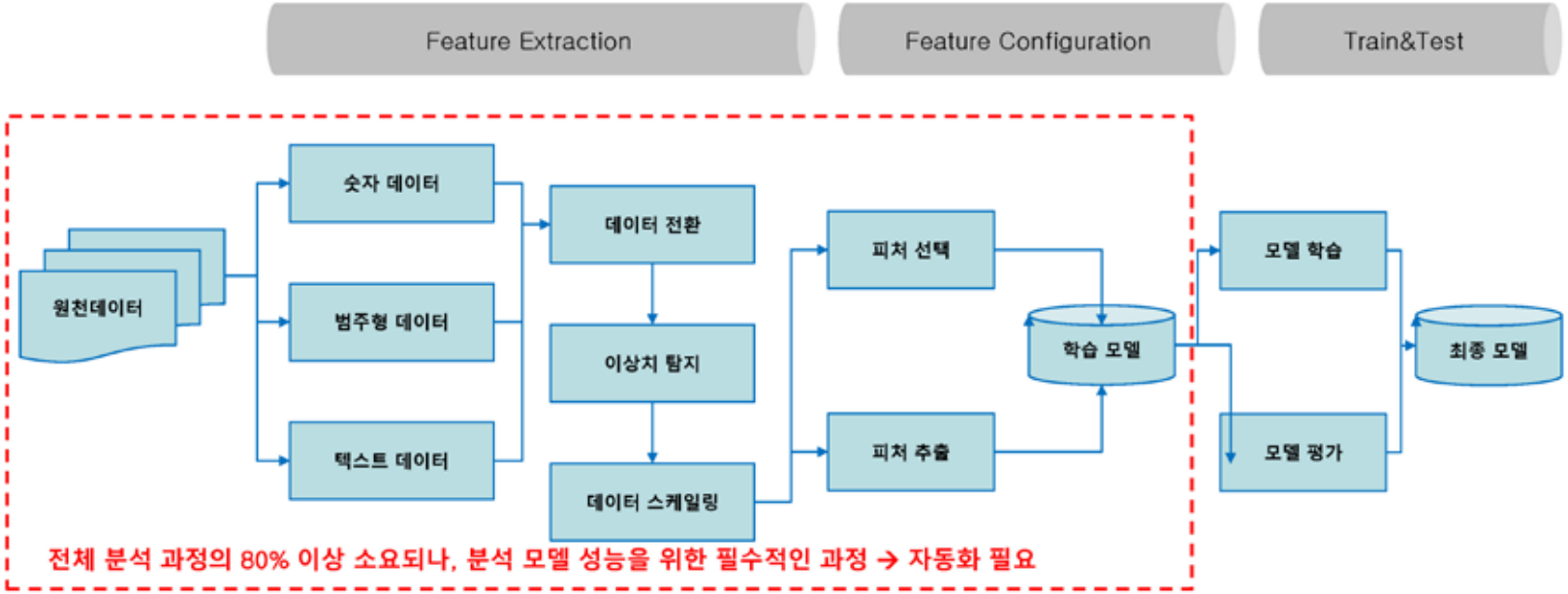


Step3. 정보별 측정 규칙 정의



4. 빅데이터 분석 자동화 구현 방안 > ② Feature Engineering

- 원천데이터로부터 피처의 이해, 개선, 구성, 평가의 단계를 수행
- 업무 전문가에 따라 2주에서 3개월까지의 기간이 소요



4. 빅데이터 분석 자동화 구현 방안 > ② Feature Engineering

➢ **Data Feature 개선을 위한 자동화된 분석 기법 적용**

데이터 변환	데이터 클린징	특성 스케일링
• 연속형 → 00100 • 범주형 → [(0,1),(2,3)] • 텍스트형 → [(0,1),(2,3)] • 이지형 → [(0,1),(2,3)]	• 중복값 제거/대체 • 누락값 제거/대체 • 이상값 제거/대체	• Standrad Scale • MinMax Scale • Robust Scale • MaxAbs Scale • Normalize

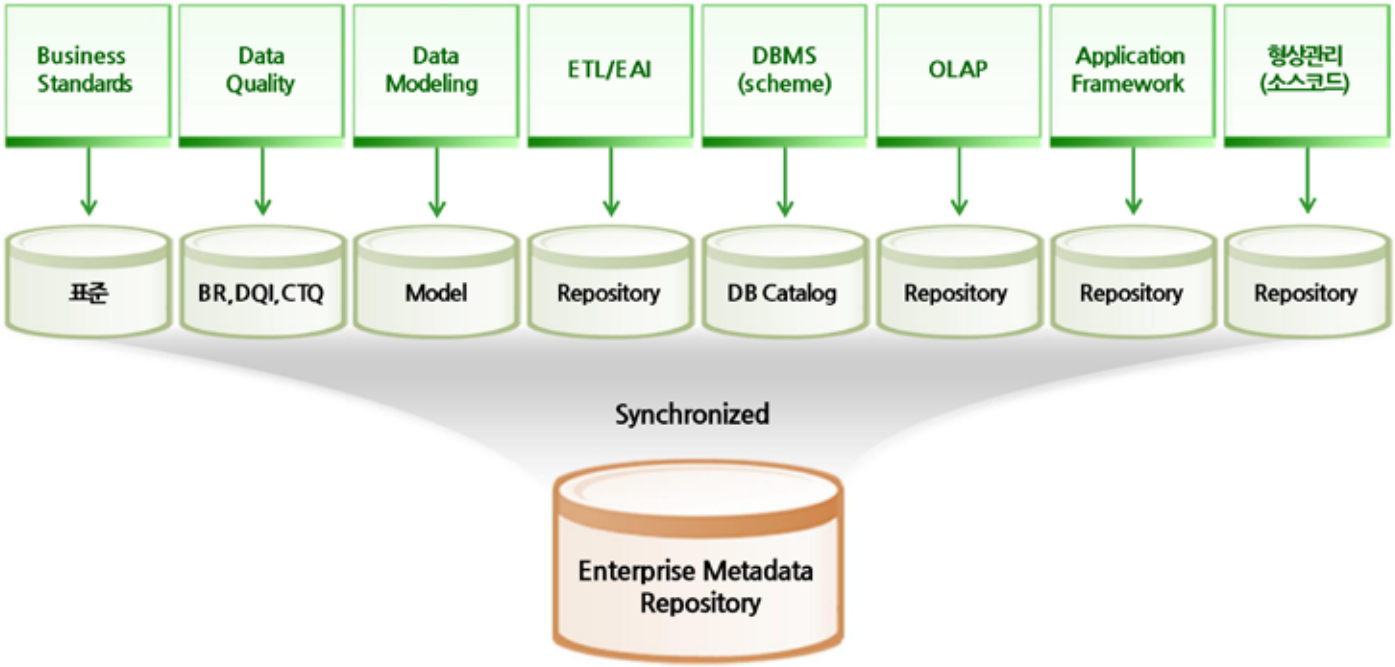
PCA

ID	한국	영국	미국	중국	캐나	프랑스	러시아
1	0	1	0	0	0	0	0
2	1	0	0	0	0	0	0
3	1	0	0	0	0	0	0
4	0	0	1	0	0	0	0
5	0	0	0	1	0	0	0

- 모델 학습에 활용 가능한 형태로 변환
- ex) 원-핫 인코딩/레이블 인코딩

4. 빅데이터 분석 자동화 구현 방안 > ③ 메타데이터 관리

➢ 데이터의 구조적 복장성 및 의미의 복잡성으로 인해 부분적인 통합관리 수준에서 전사 차원의 넓이와 Business end-user 레벨 깊이까지 다양한 IT 환경의 메타데이터 유형이 통합된 메타데이터 영역으로 확대



4. 빅데이터 분석 자동화 구현 방안 > ③ 메타데이터 관리

➢ 다양한 메타데이터는 궁극적으로 복잡한 메타데이터 구조를 단일 사용자 관점에서 통일된 View를 제공하여야 함

유형(Type)		목적(Purpose)	설명(Examples)	사용자(Audience)	원천(Origin)
업무(Business)		비즈니스 측면에서 데이터의 의미 이해를 돕기 위함	각종 문서, 보고서, 사용자 화면에 나타나는 업무용어 등	현업사용자	매뉴얼 (Manual)
기술 (Technical)	Development	기술적 구성요소 간의 상호참조 및 연결구조의 이해를 돕기 위함	시스템 정보, DB 테이블 정의, 구조, Size, 테이블간의 관계, 속성 정의, 유형, Size, Source, Target, Mapping	개발자	시스템 이미지정보 System Captured
	Operational	운영시스템의 모델 및 품질관리지원 데이터웨어하우스 운영처리를 지원		운영담당자 DW 관리자	시스템 생성정보 System Generated
	Relationship	모든 가능한 Object에 대한 연관성 추적 분석을 지원	Application, Component, Document, DB, Message, Log	모든 사용자	매뉴얼 / Agent 등

4. 빅데이터 분석 자동화 구현 방안 > ④ EDA

➢ 탐색적자료분석은 데이터의 특징과 내재하는 구조적인 관계를 알아내기 위한 분석 기법으로 탐색 과정을 통해 얻은 통계모형/시각화 등을 바탕으로 미지의 특성을 파악하는 증거 수집의 과정

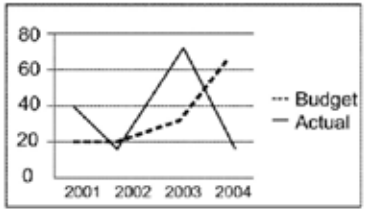
자료의 일부가 파손되었을 때 영향을 적게 받는 성질

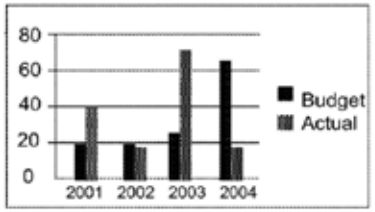
데이터의 또 다른 특징의 발견

자료를 시각화하여 표현

분포의 대칭성, 선형성, 분산의 안정성, 관련변순의 가법성 등이 나타나도록 자료를 재표현

Year	Budget	Actual
2001	20	38
2002	19	16
2003	28	71
2004	64	16

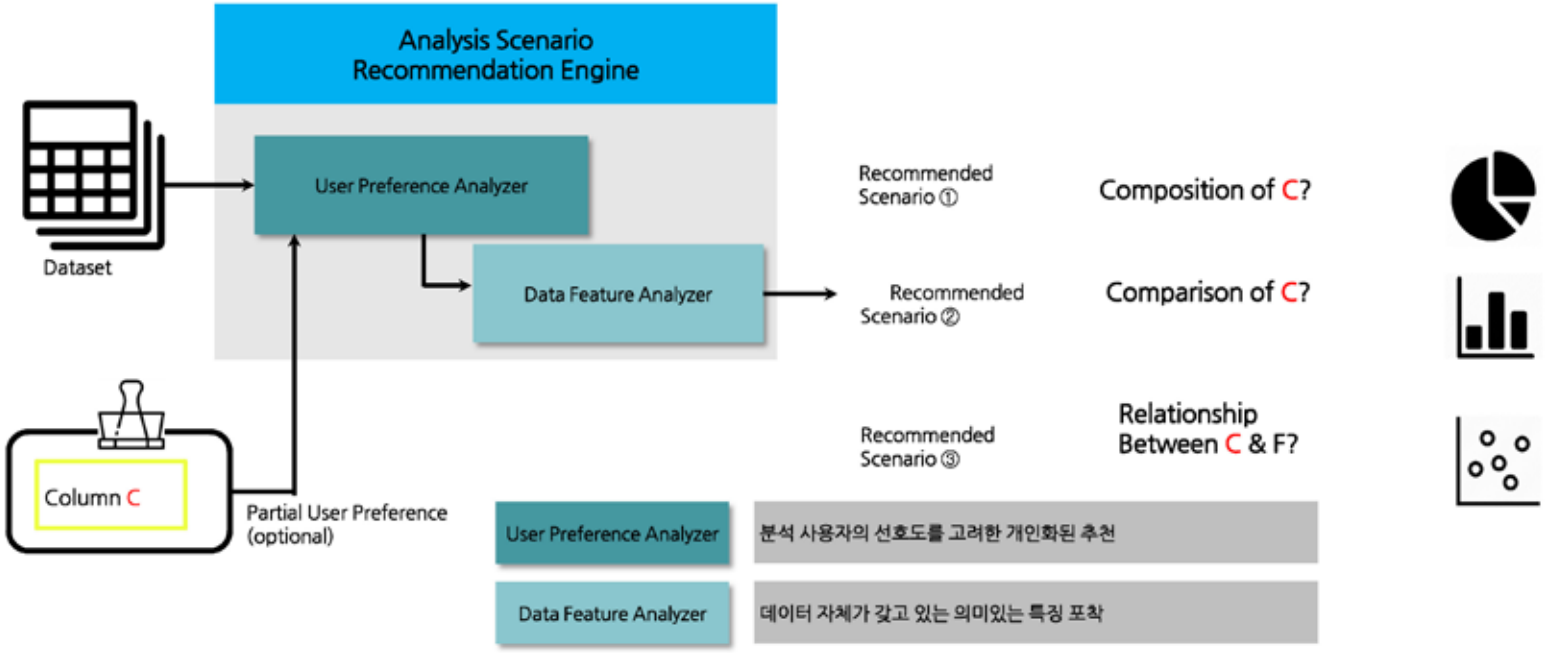




모수 데이터에 대한 표본 추출을 수행하고, 이에 대한 통계적 검정방법을 활용

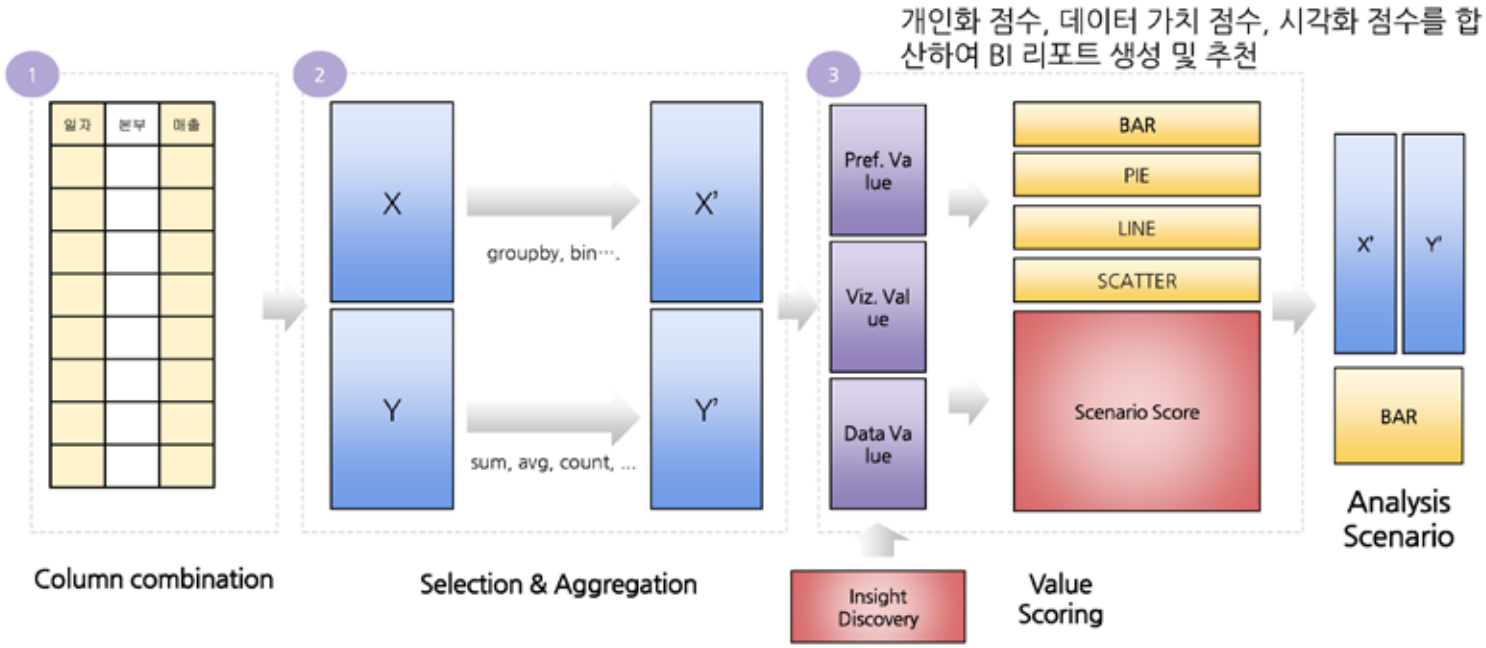
4. 빅데이터 분석 자동화 구현 방안 > ⑤ 분석 시나리오 추천

➢ EDA 과정을 수행하며 데이터에 대한 가설을 떠올리는데, 해당 과정을 자동화하여 Insight가 되는 리포트를 추천



4. 빅데이터 분석 자동화 구현 방안 > ⑤ 분석 시나리오 추천

➢ 데이터의 의미있는 정보를 추출하여 자동적으로 BI 리포팅을 생성하여 사용자의 원활한 탐색적 데이터 분석 과정을 지원



4. 빅데이터 분석 자동화 구현 방안 > ⑤ 분석 시나리오 추천

➢ Key-Cause Analysis란 사용자가 분석 결과물을 다양한 관점으로 이해하여 인사이트를 도출할 수 있는 주요 원인 설명 및 시각화를 제공

분석 자동화를 통한 다양한 관점의 데이터 해석

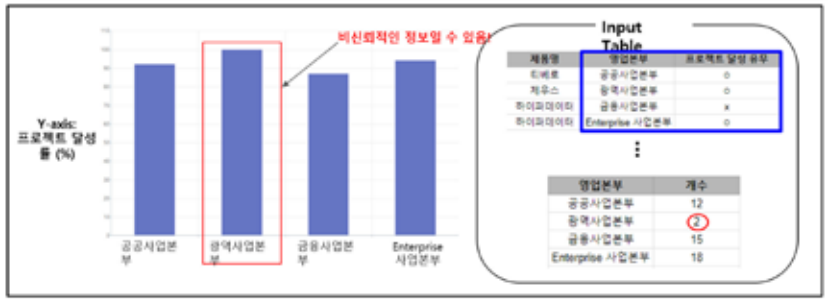
신뢰(Reliability)
분석 결과물의 비이상적 특이점 도출

상관관계(Correlation)
다른 Attribute 들과의 상관관계 파악

경향(Tendency)
시계열, 특정 조건에 따른 경향성 파악

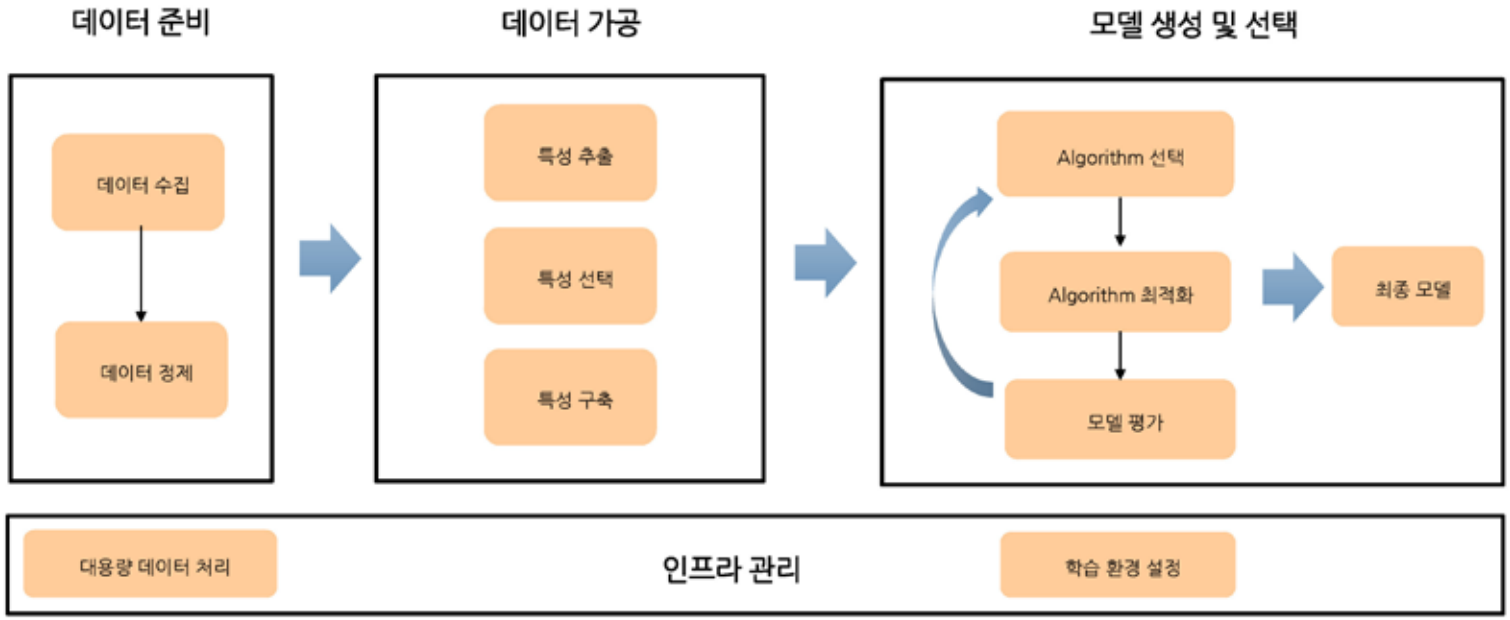
분포(Distribution)
특정 상황에서 Attribute 들간의 분포 파악

분석 결과물에서 사용자가 선택한 특정 Target와
관련된 Explanation 시각화 및 구체적인 설명 제공



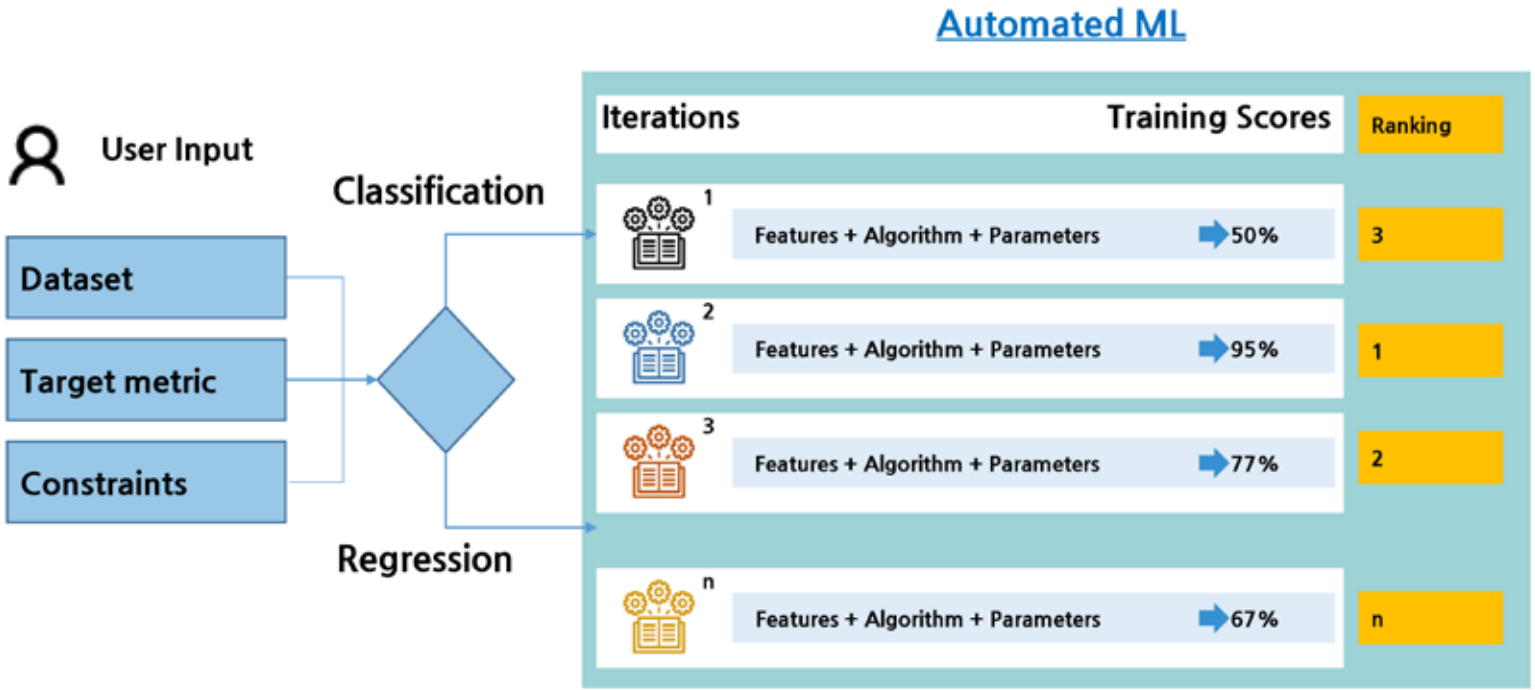
4. 빅데이터 분석 자동화 구현 방안 > ⑥ AutoML

➢ AutoML은 원시 데이터에서 배포가능한 기계학습 모델에 이르는 전체 파이프라인을 다루며, 비전문가도 손쉽게 인공지능 기술을 활용할 수 있음



4. 빅데이터 분석 자동화 구현 방안 > ⑥ AutoML

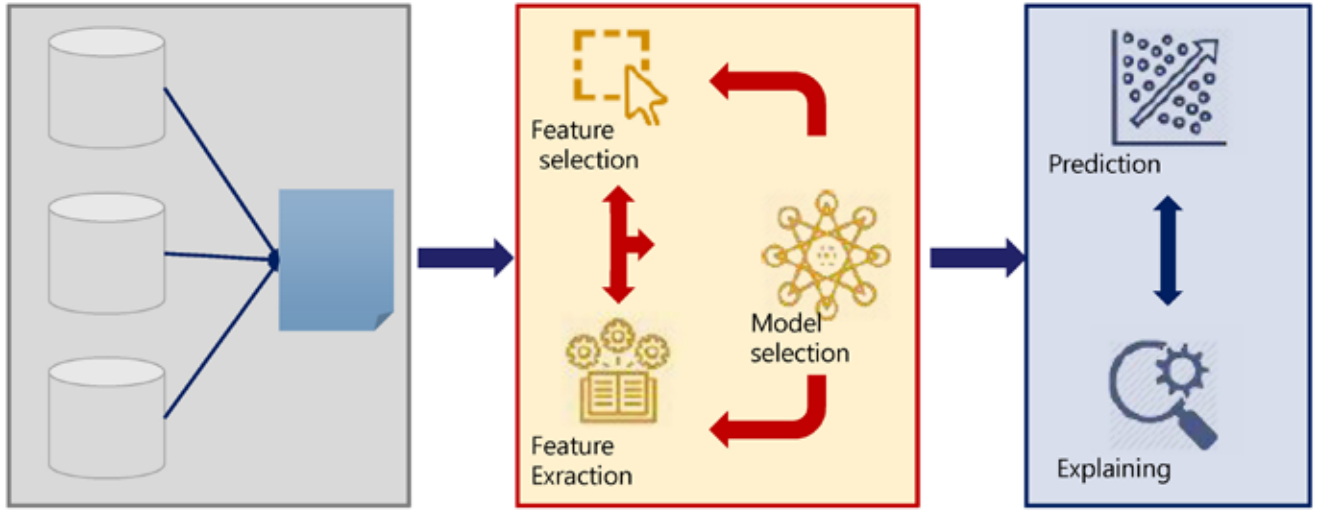
➢ 실제 문제에서 기계학습을 적용하는 프로세스를 자동화



5. TmaxBI HyperData

수정 예정

"비전문가도 손쉽게 데이터 분석이 가능한 **스스로 학습하고 결과를 설명하는 Automated ML 플랫폼** "



5. TmaxBI HyperData > 특징 및 차별성

1 Data Virtualization

데이터의 실시간성 분석

- 다양한 데이터 유형의 통합
- 데이터의 실시간성 분석을 통한 민첩한 비즈니스 대처
- 데이터의 물리 복제 최소화

2 Cloud BI

자유로운 유연성을 갖춘 데이터 탐색 및 분석

- 데이터 분석 통합 개발 환경 제공
- Flow 기반의 데이터 처리를 통한 손쉬운 데이터 활용
- 다양한 데이터 시각화 및 탐색

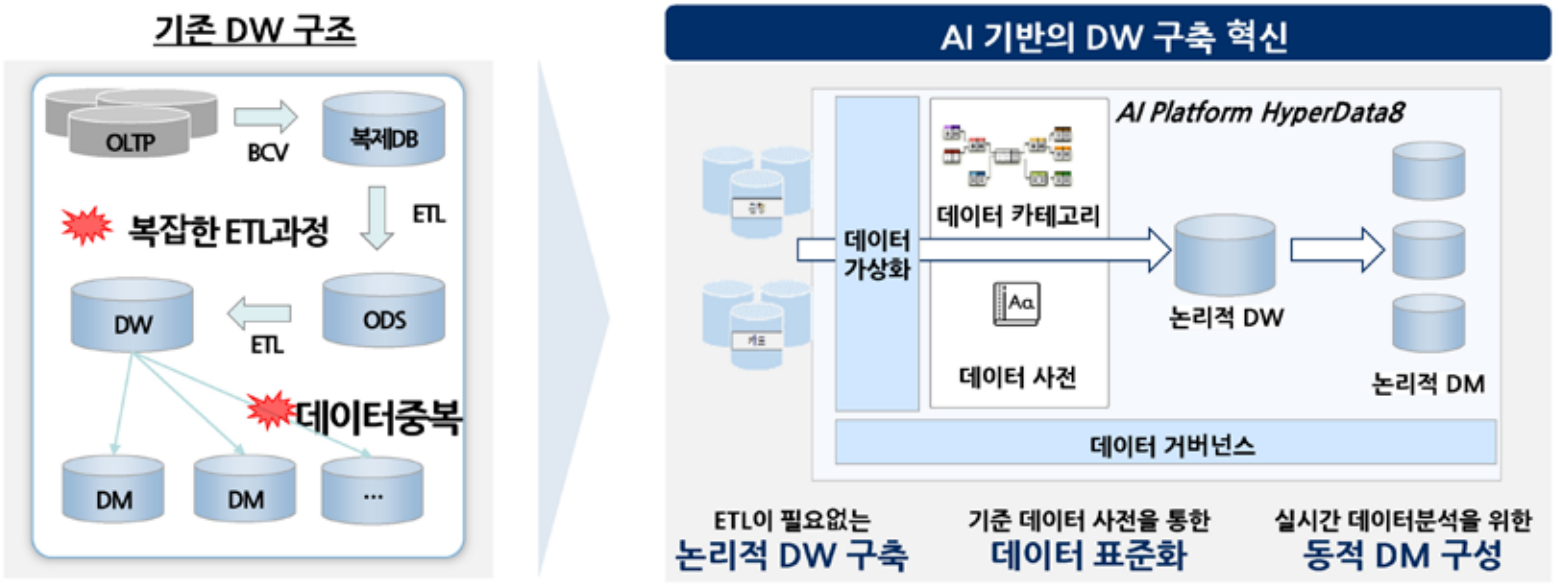
3 AutoML XAI

사용자 중심의 AI 분석 편의성 제공

- EDA 자동화
- 머신러닝 알고리즘 자동 추천
- AI 블랙박스 문제 해결을 위한 설명가능한 AI

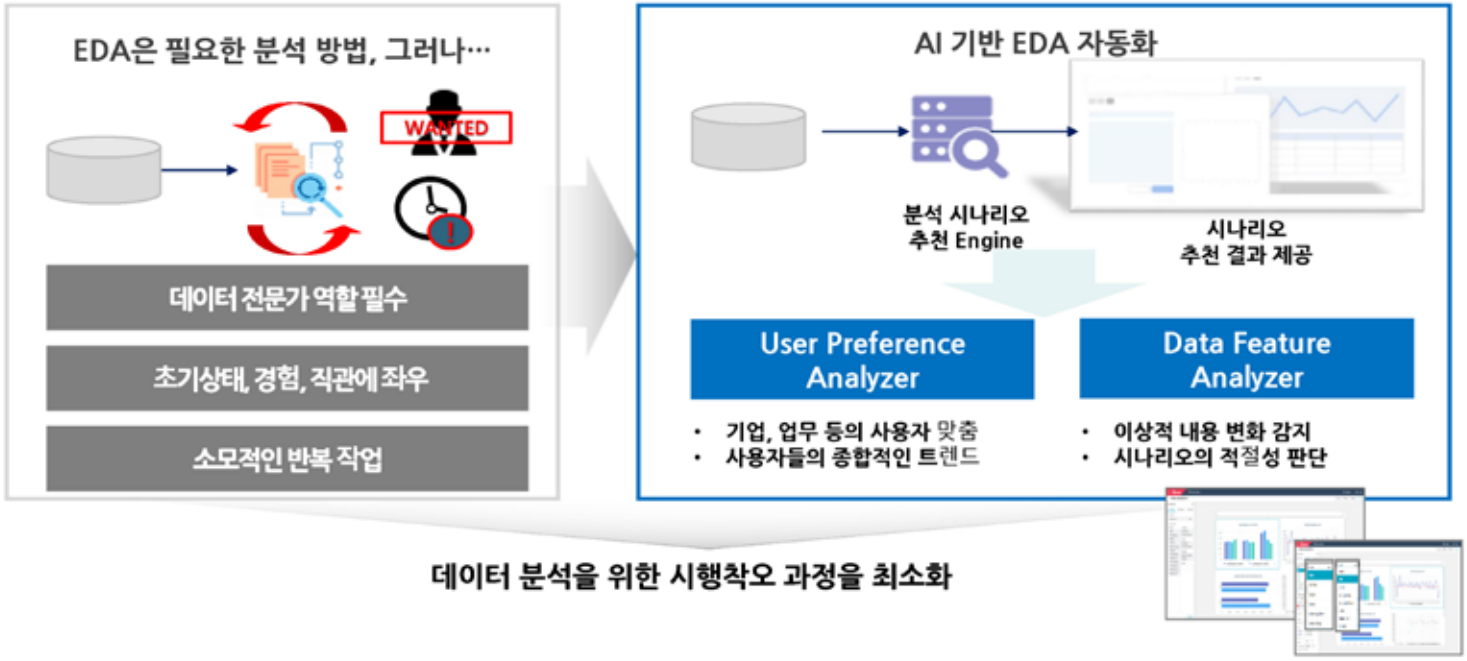
5. TmaxBI HyperData > 데이터 가상화

메타데이터, 데이터 카탈로그 기반의 Logical DW → 손쉬운 데이터 획득

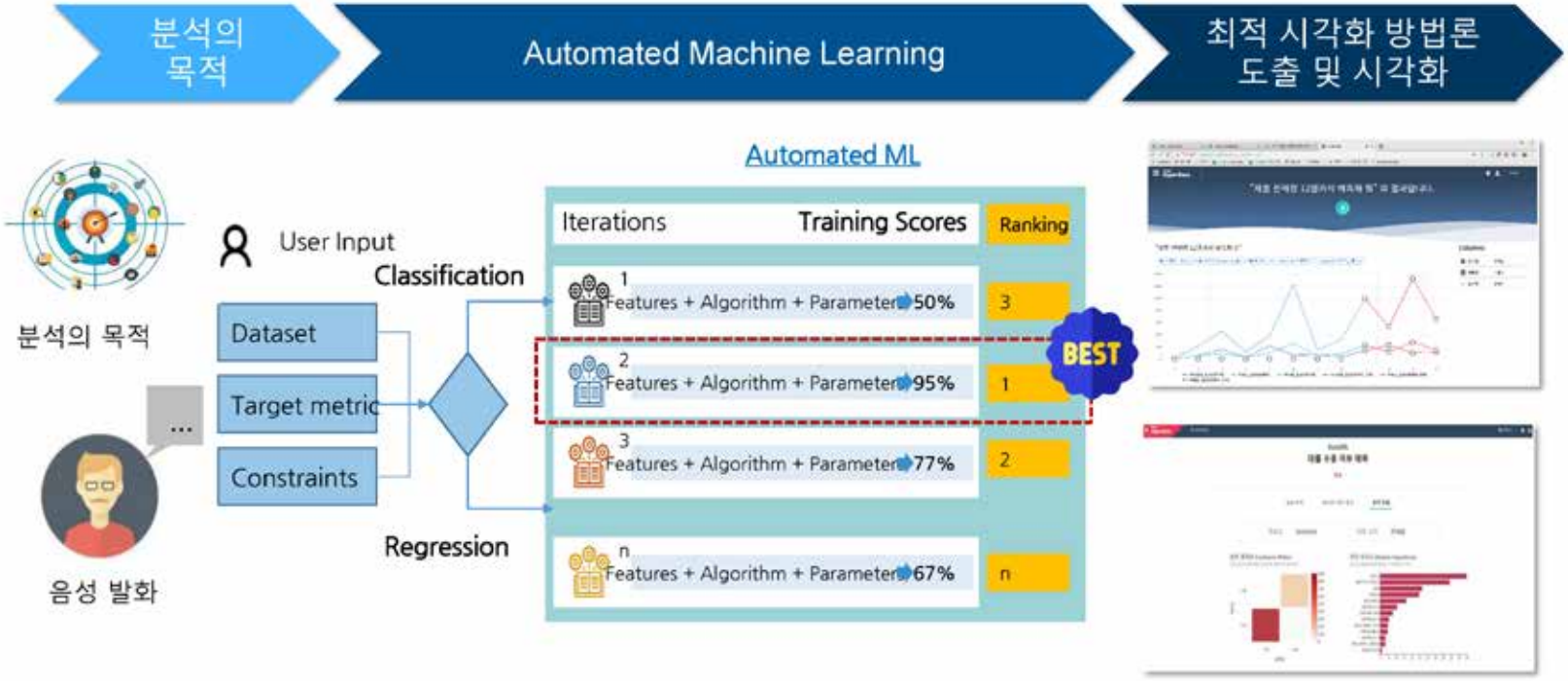


5. TmaxBI HyperData > Cloud BI

분석 과정의 자동화와 지능화로 데이터 기반 인사이트 제공

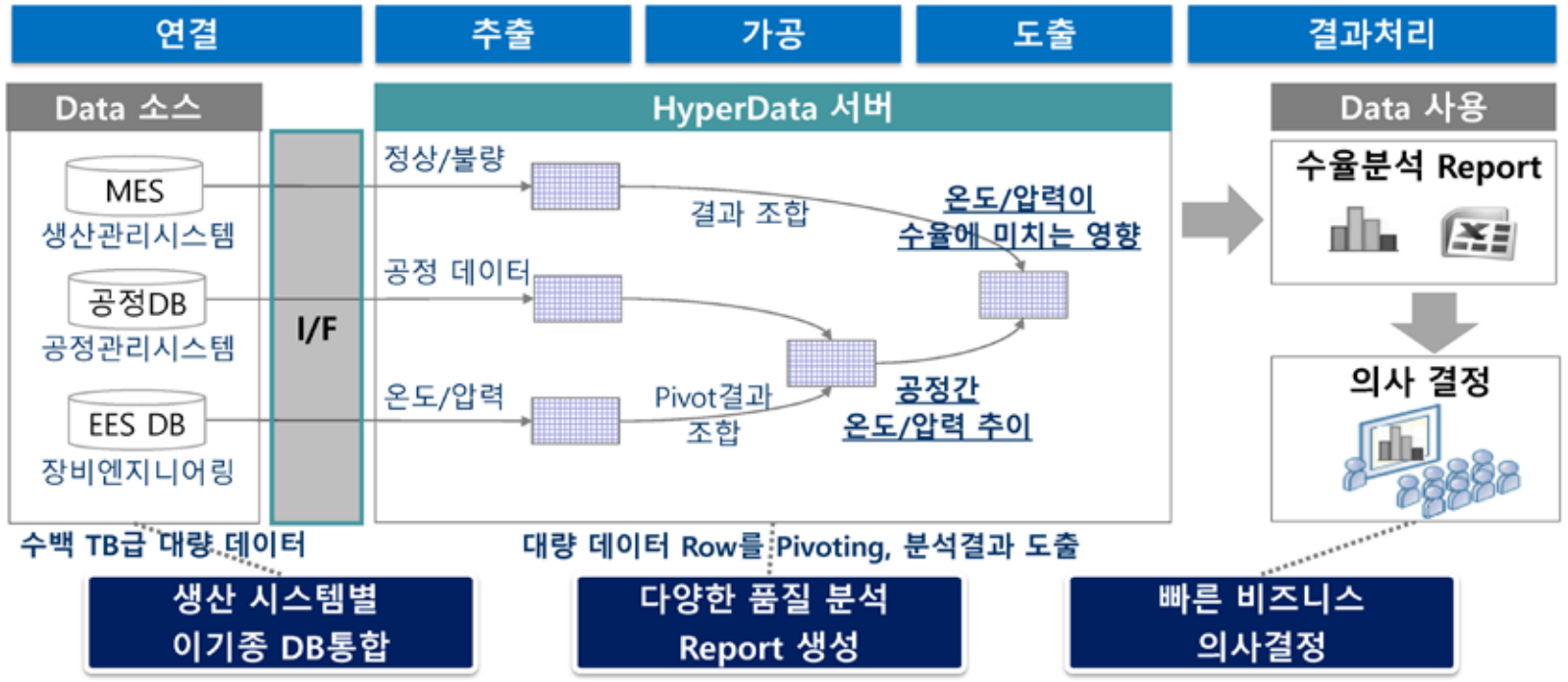


5. TmaxBI HyperData > AutoML&XAI



5. TmaxBI HyperData > 도입 사례

➢ Self Service 사례 – 사용자가 이기종 데이터를 추출하여 통합 분석이 가능한 Self Service BI 환경



5. TmaxBI HyperData > 적용 사례

➢ 금융권 대출

	Tmax HyperData	MS Azure	H20 Driverless AI	Kaggle / OpenML
MiniBooNE	94.9871	94.6870	94.6765	94.8000
Credit-g	77.0000	77.8989	78.3736	78.6000
은행 고객 정보	98.9000	98.7600	98.7400	98.3000
EEG eye state	95.2603	93.1709	94.8992	98.5200
Adult Census Income	87.3177	88.517	87.5625	86.2000

정확도(%)

Better Cloud, Better Tomorrow

감사합니다.

Tmax