

거대 추천 모델을 통한 SKT 마케팅 혁신 프로젝트

Big Model (“Large Scale AI”) 기반 추천 혁신

AIX / Data R&D / Customer Analytics팀

‘23. 3

Contents

- 1. Why**
- 2. Now**
- 3. Future**

Contents

1. Why

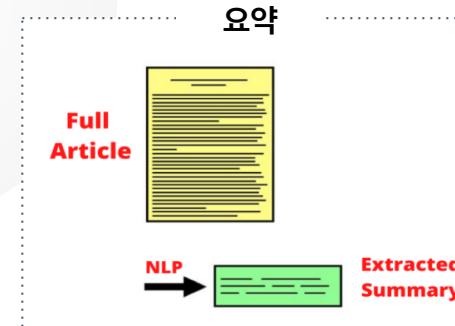
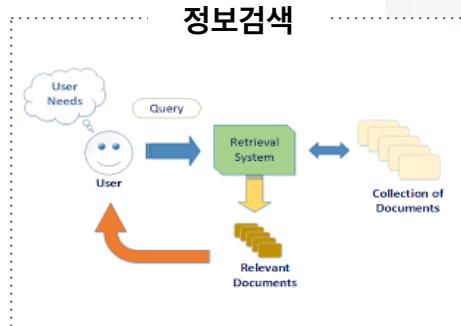
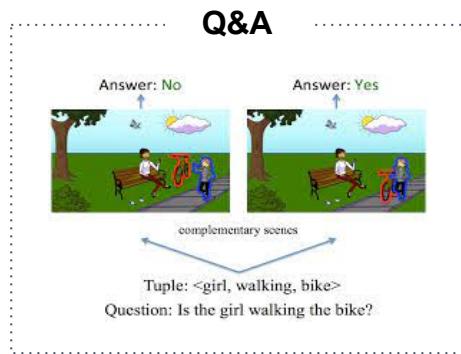
2. Now

3. Future

패러다임의 변화:

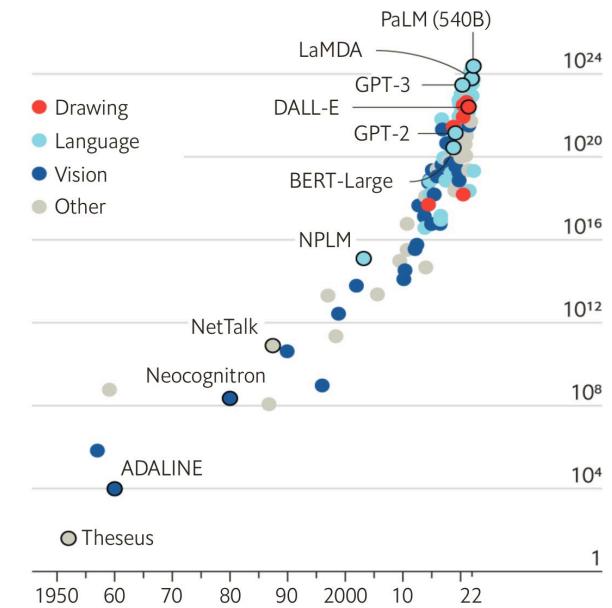
방대한 데이터로 거대 모델(Large Scale AI) 학습 시 개별 모델 대비 압도적인 성능과 편의성, 범용성을 가짐
이는 자연어, 이미지 뿐 아니라 추천/타겟팅 영역에서도 지속적으로 시도되고 있음

초거대 언어모델



The blessings of scale

AI training runs, estimated computing resources used
Floating-point operations, selected systems, by type, log scale

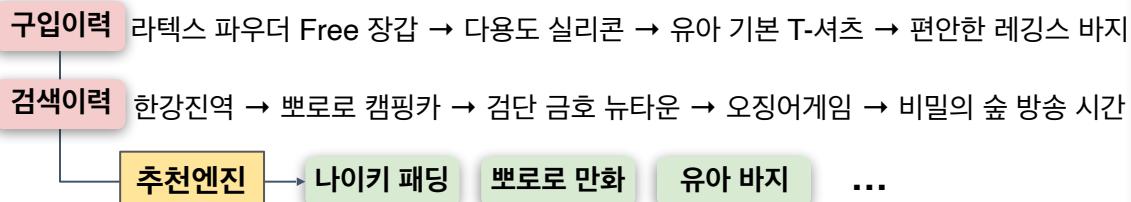


Source: Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[국내] 네이버, 카카오 등 국내 기업들은 대량의 자사 고객 데이터를 이용한 거대 추천 모델을 구축했으며 이를 이용해 다양한 채널에서 추천, 타겟팅을 진행하고 있음

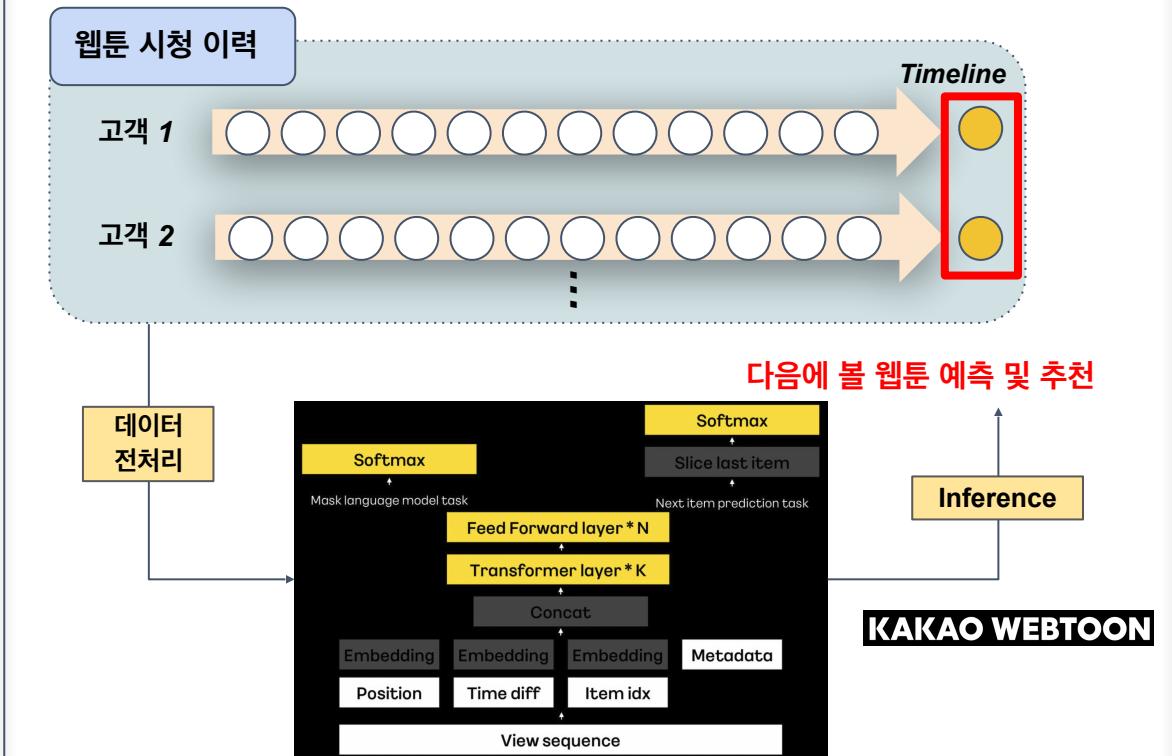
*네이버: 언어 모델 기반 거대 추천 모델 개발 (GPT)

고객 검색쿼리, 구입 상품 단어를 이용하여 Transformer 기반 모델 통한 pre-trained user representation 모델 구축



카카오: 딥러닝 기반 추천 모델 개발 (BERT)

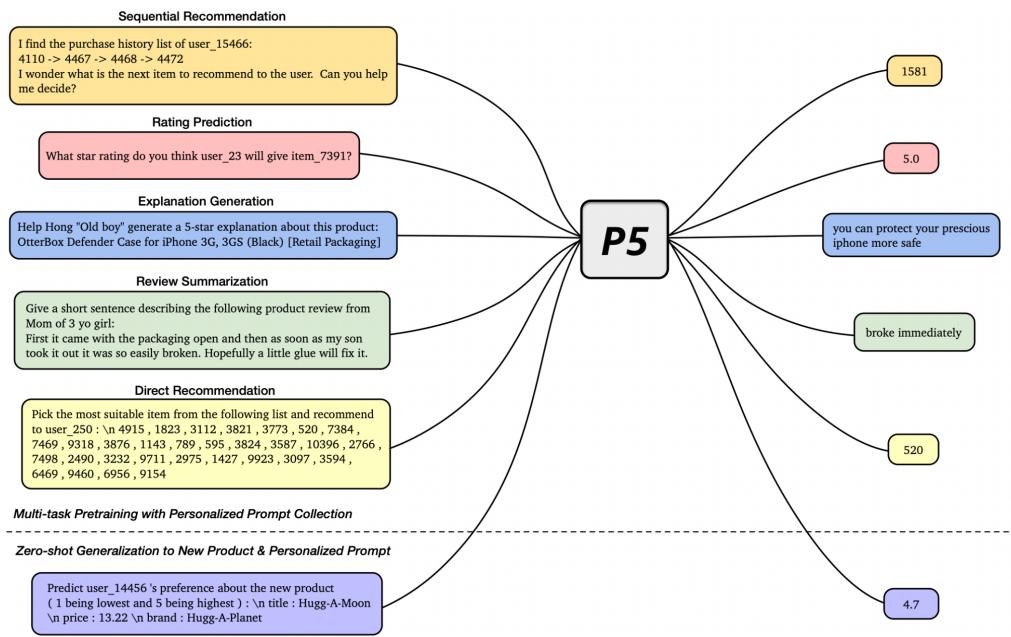
BERT기반 user representation 모델 구축 및 카카오웹툰 적용



[국외] 구글, 메타, 엔비디아 등 해외 빅테크 기업들 역시 초기대 추천 모델을 구축하여 이용 중임

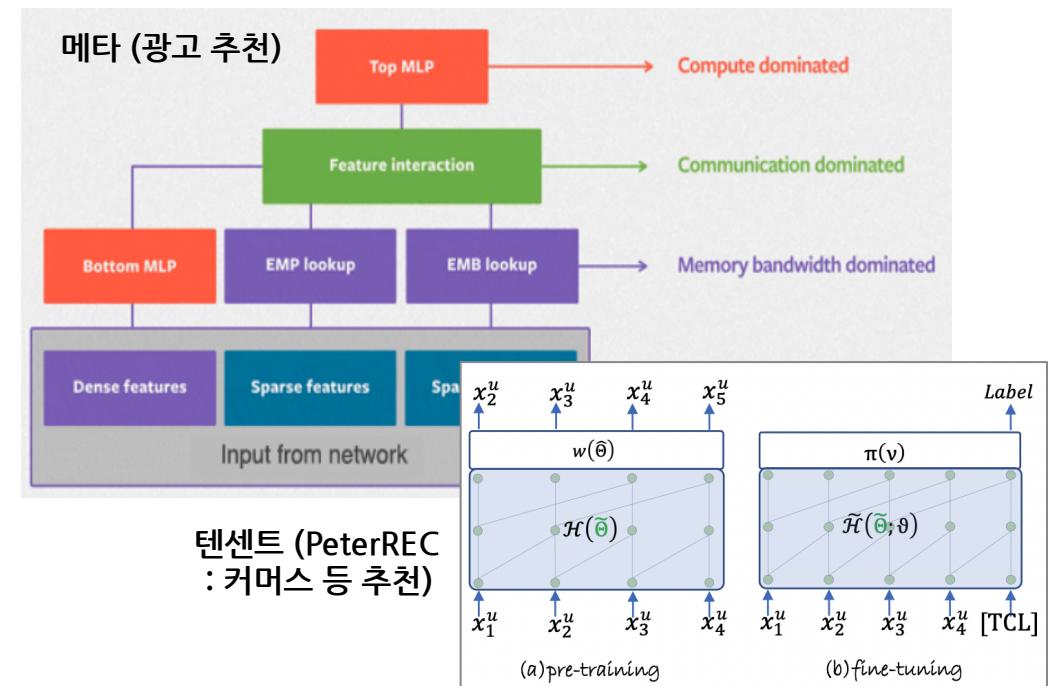
구글 : Recommendation AI 서비스

GCP에서 구글의 여러 추천 모델 사용할 수 있도록 서비스 하고 있음. 한편, 구글 T5 거대 언어 모델의 framework와 동일한 자연어 기반의 추천엔진 역시 학계에서 개발



메타/텐센트 : 딥러닝 기반 추천시스템

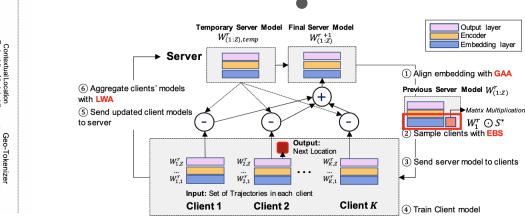
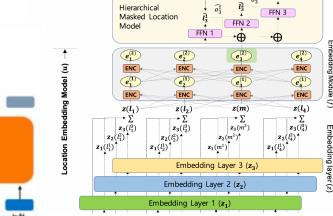
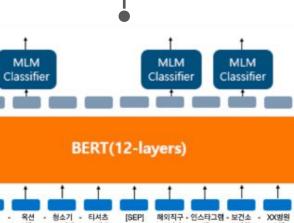
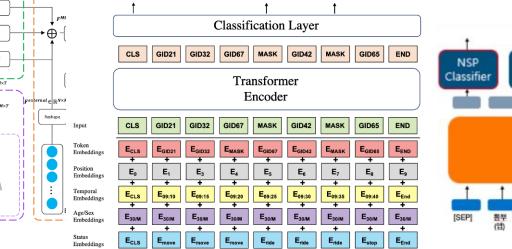
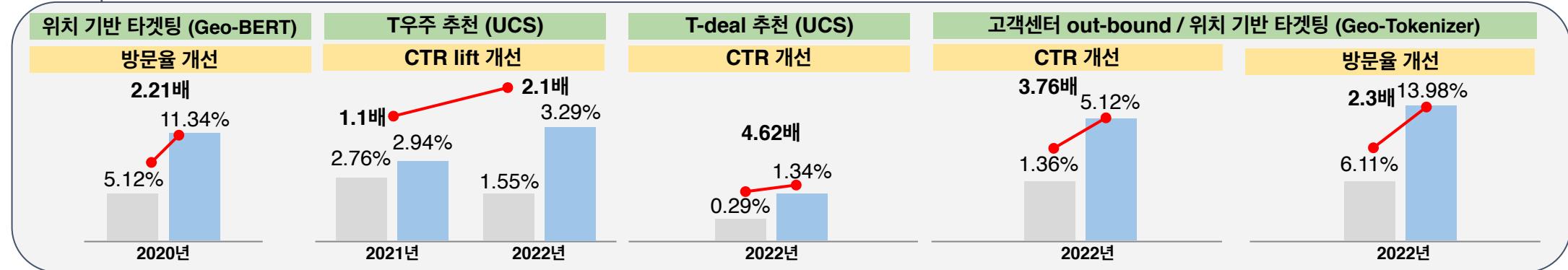
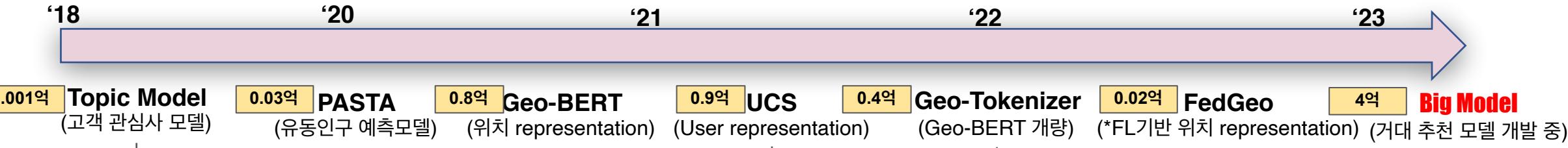
사용자 페이스북 로그 데이터 활용한 딥러닝 추천 시스템 개발/사용 중
텐센트는 자사 앱내 로그를 토대로 대규모 추천 모델 구축 및 연구



당사 거대 모델 개발 현황



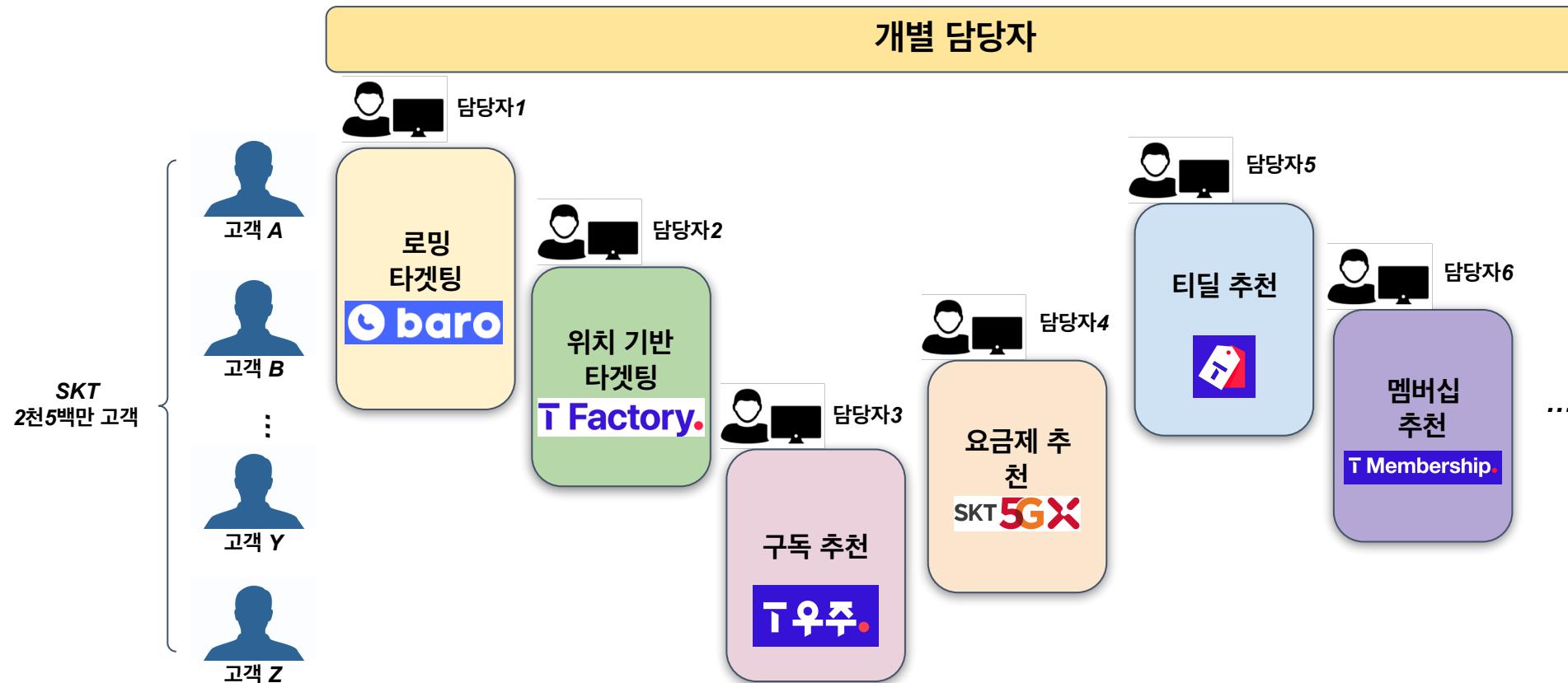
GPT, BERT 등 Language Model 기반 추천, 타겟팅 거대 모델을
‘18년부터 개발해 왔으며 논문 게재 및 특허 출원 지속해 옴



* Federated Learning

개별 담당자가 각 모델을 서로 다른 데이터셋으로 운영하여
모델 운영 효율성이 낮고 전사적 관점에서의 통합 관리 (Fatigue 관리 등) 가 어려움

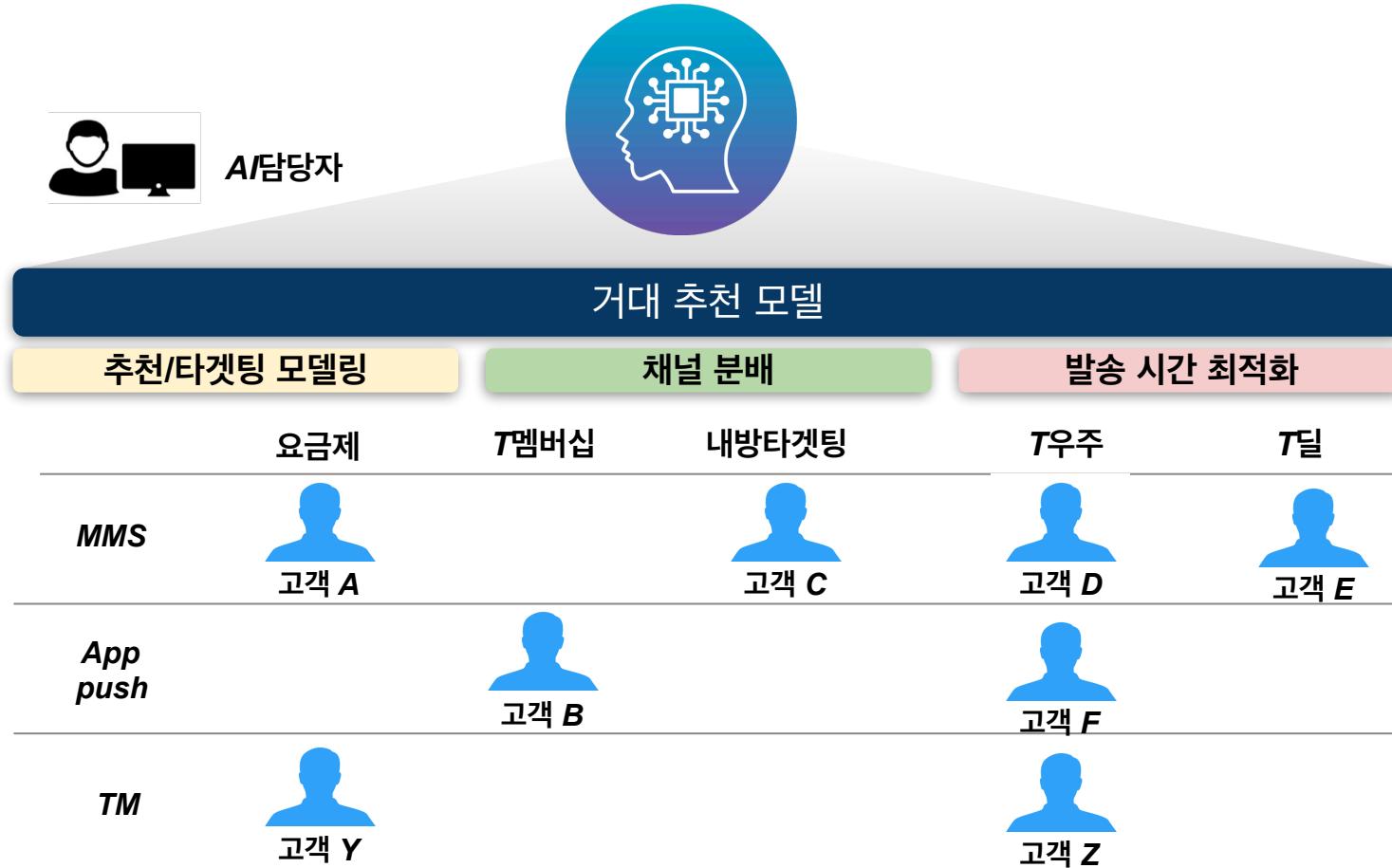
“특정 고객이 어떤 채널의 어떤 메시지/push를 받았는지 통합적 확인 불가”



거대 추천 모델 필요성



방대한 행동로그 데이터를 활용한 거대 추천 엔진을 토대로
전사의 모든 추천/타겟팅의 성능 향상과 운영 효율성 제고 및 거버넌스 일원화



chatGPT를 “거대 추천 모델”로 활용할 수 있을까?



chatGPT 자체를 당사 추천에 fully, 직접적으로 적용 시
(1) 비용, (2) 운영, (3) 민감정보 이슈 존재

고비용/저효율

- [고비용] Azure OpenAI 요금 체계는 사용성과 관계없이 시간당 비용 발생

* 학습 시간에 따른 비용
* 호스팅 시간에 따른 비용
* 토큰 입력 당 비용 (GPT-4 3.5대비 15배)

- [저효율] 채널별, 주기적 chatGPT 미세 조정 학습 비효율적

* 고객 feedback이 매일 인입되며 이를 모델에 반영하기 위한 지속적 미세조정 학습 필요
* T-우주, T-deal, T-membership 등 당사 보유 채널 별로 미세조정 학습 및 추론 필요

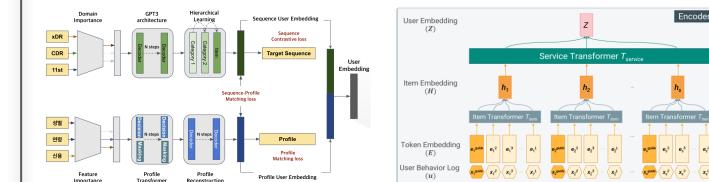
운영 이슈

- [운영이슈] 타사 플랫폼에 fully 종속될 경우 운영 상 다양한 이슈 대응이 어려움

* 기존 GCP autoML 운영 → 현업 feeding 위한 조건 (동시간 대 추론 가능한 고객 수, 학습 모델 수 등) 제한 존재
* 모델 customizing 제한 (기술내재화 불가)

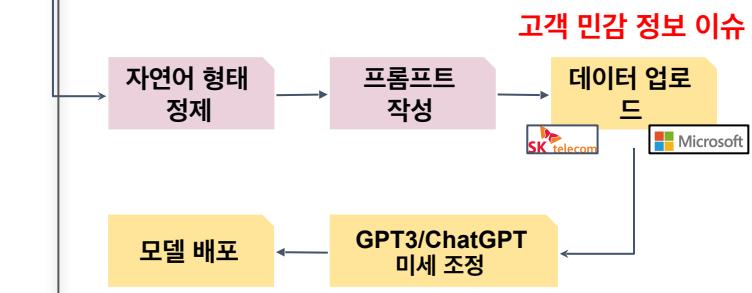
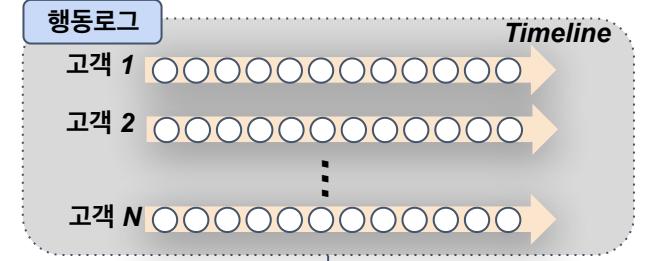
- [기술경쟁력] chatGPT가 커버 못하는 부분 관련 전문성 가지는 모델 개발 필요

- (참고) 네이버 사례
자체 추천 엔진 개발 (**CLUE**) → 다양한 채널 내 자유로운 운영 및 기술 내재화



민감정보 risk

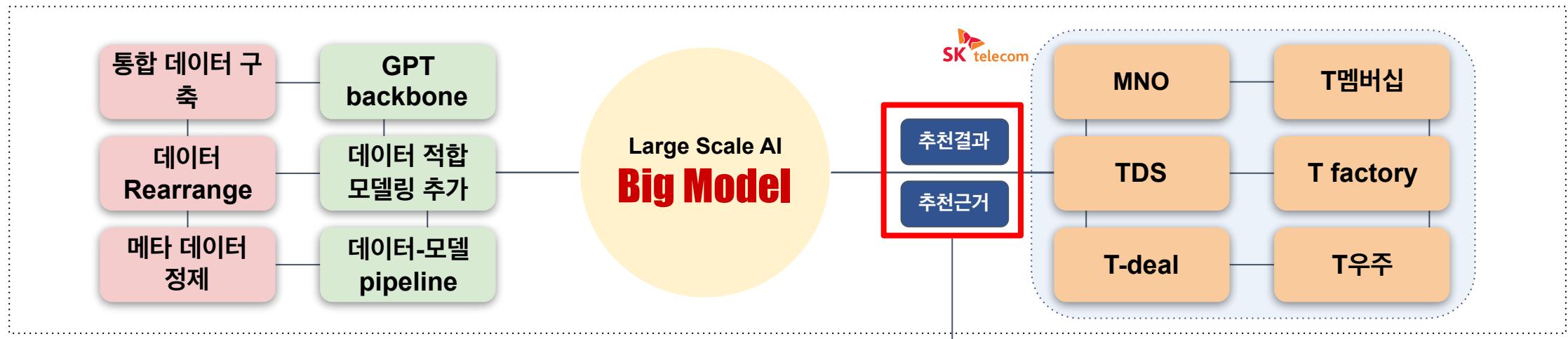
- 고객 정보 (앱/위치 등) 를 이용한 Azure OpenAI 에서의 모델 튜닝은 정보보호 이슈 존재



“거대 추천 모델”의 자체 개발 필요성



우리 데이터의 특성을 가장 잘 반영할 수 있는 거대 추천 모델(Big Model) 개발을 통해 비용 절감, 운영 효율화, 개인정보 보호 및 기술 내재화를 도모 (+ chatGPT 적용 효과 극대화)



On-Premise 내 거대추천엔진 학습/추론/업데이트 비용 절감

- 지속적으로 쌓이는 고객 데이터 수시로 업데이트 → 고객 특성 현행화



운영 dependency 최소화

- 내부 엔진으로 고객 특성을 최적으로 반영한 모델 구축하고 기능 추가나 실제 배포 상의 다양한 이슈를 신속하게 해결



고객 민감정보 이슈 risk 해소

- 위치, 앱/웹 사용 등 민감 정보 risk를 줄이고, 내부 모델 결과와 ChatGPT 조합으로 추천 이유/우선순위 관련 script 생성 (고객센터/유통망 지원)

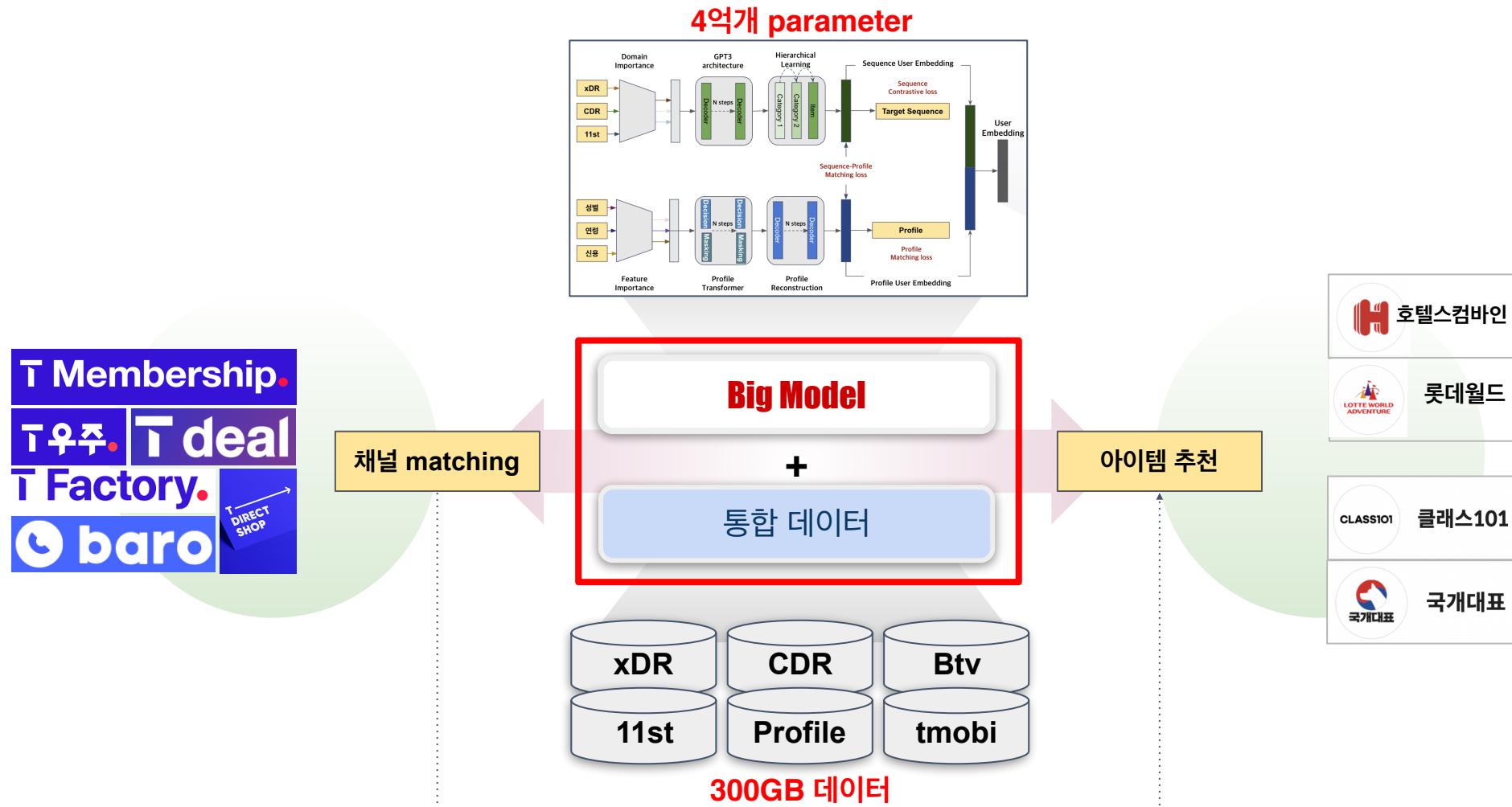
Contents

1. Why

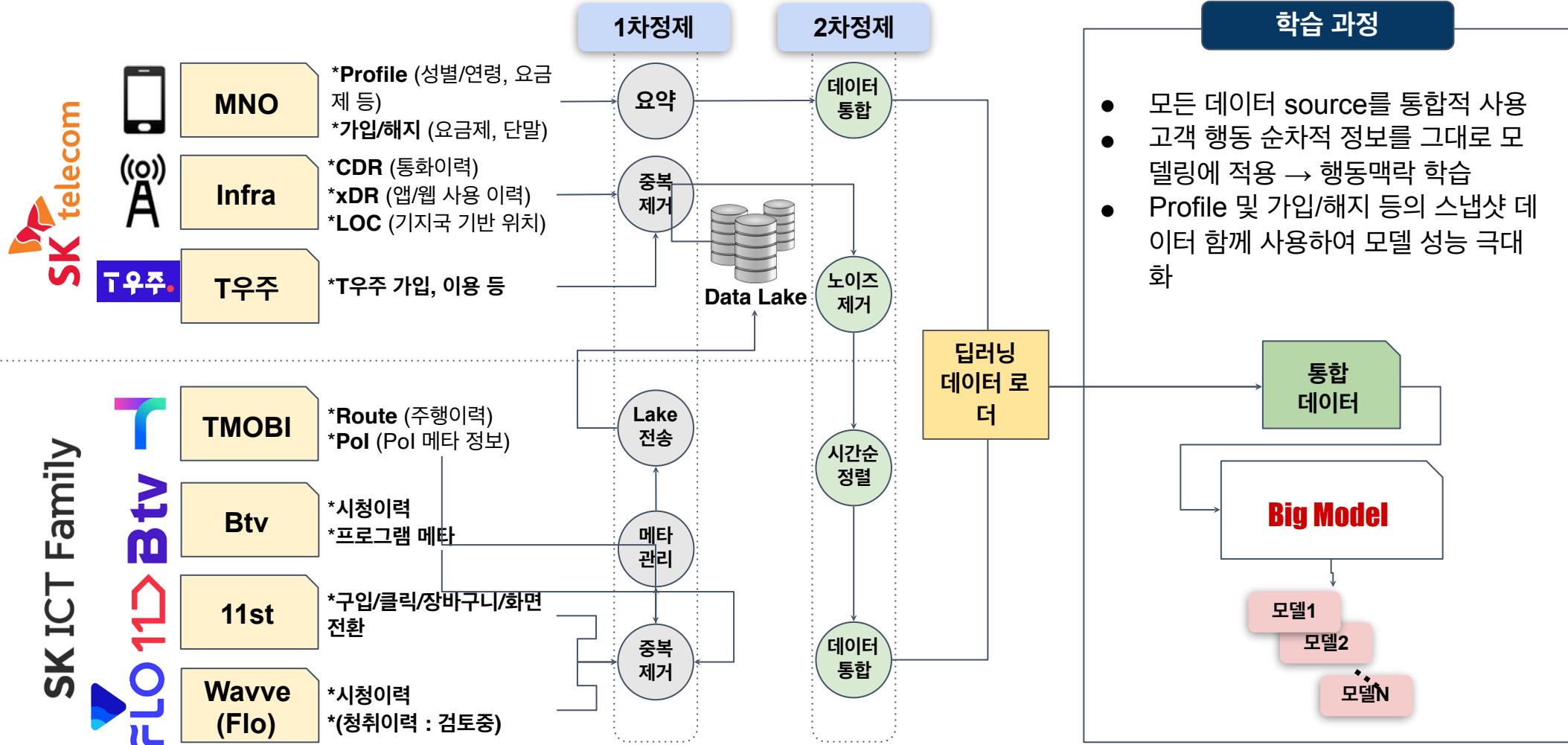
2. Now

3. Future

거대 추천 모델(Big Model) 개발을 위해 전사 데이터 정제 및 통합 적재 작업
및 proto-type 모델 구축 완료 ('23년 3월)



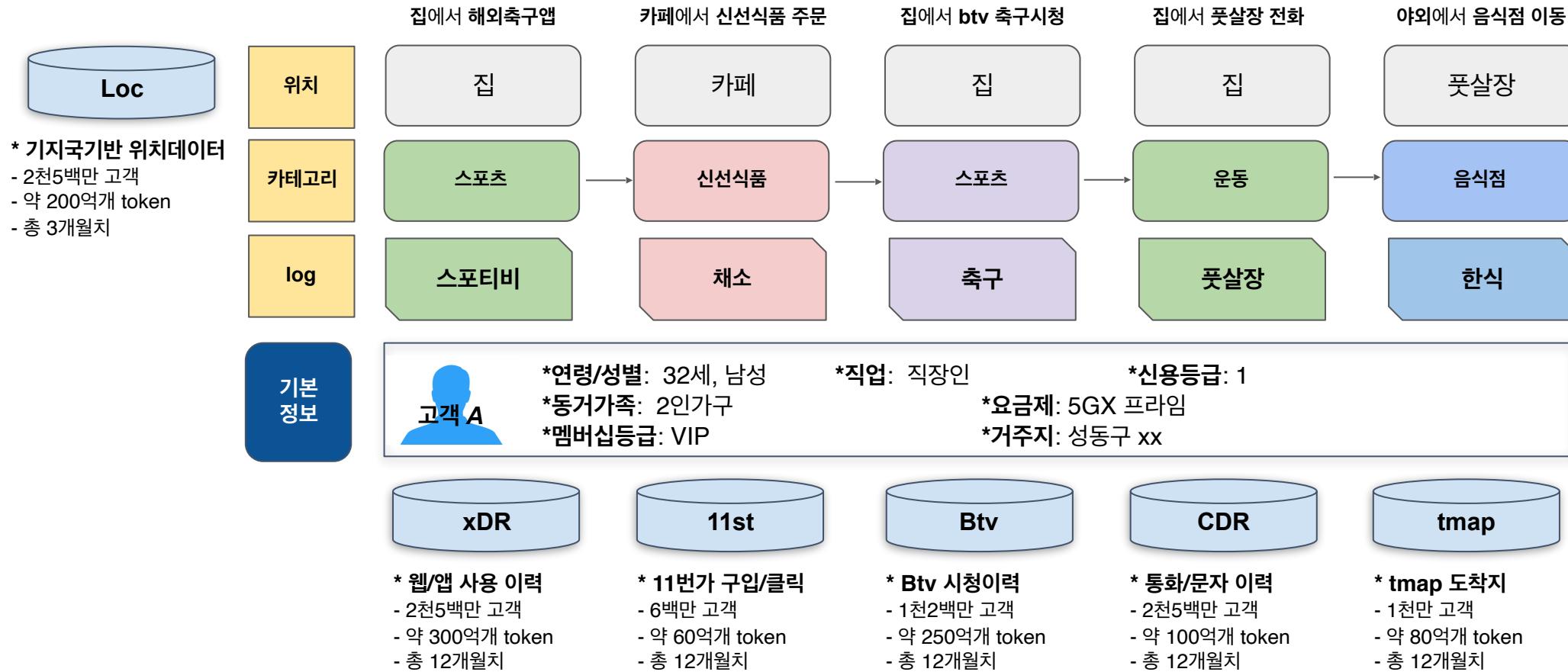
[데이터] 데이터 파이프라인



[참고] 당사 데이터 특장점



SKT는 다양한 도메인 앱 사용이력, 통화이력, Btv 시청이력 등 의 고객 행동 로그 및 profile 정보 성별, 연령, 신용등급 등 를 보유하고 있으며 이를 준실시간성 대규모 journey로 구성 가능

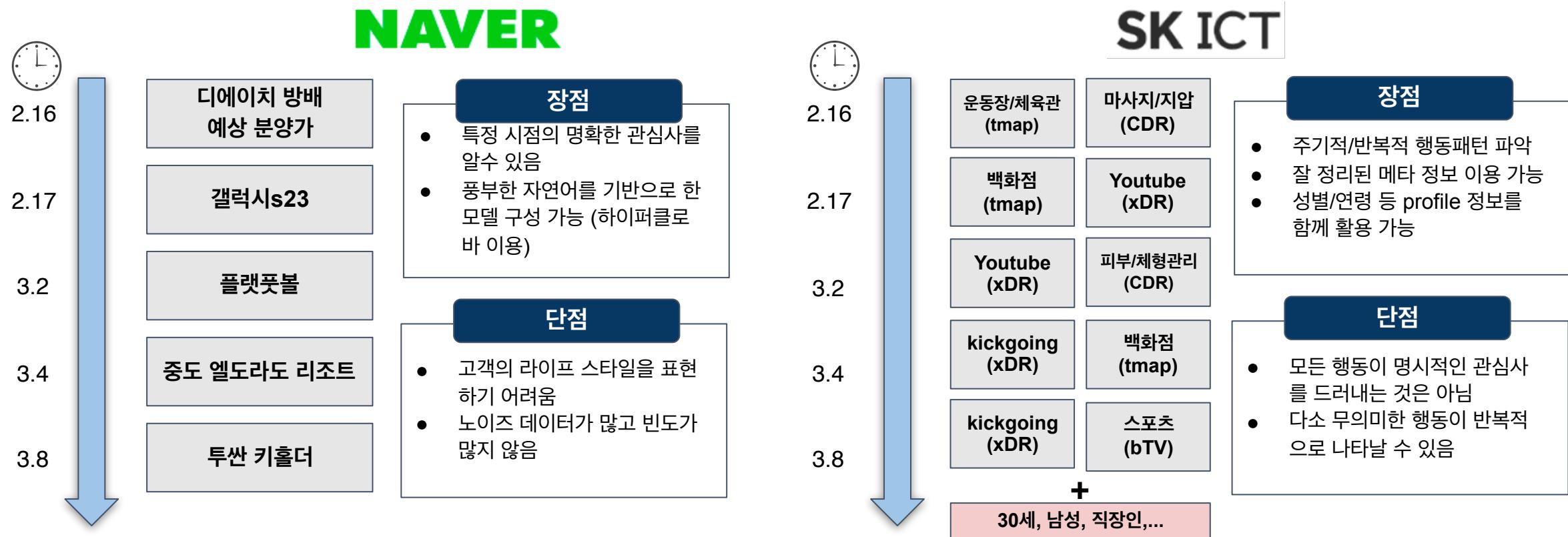


[참고] 당사 데이터 특장점



네이버의 자연어 기반 검색 쿼리 데이터와 비교하여
SKT의 행동로그는 고객의 라이프 스타일을 직접적으로 반영

당사 데이터는 고객의 주기적/반복적인 행동 패턴은 통합적인 관심사를 표현하기에 적합

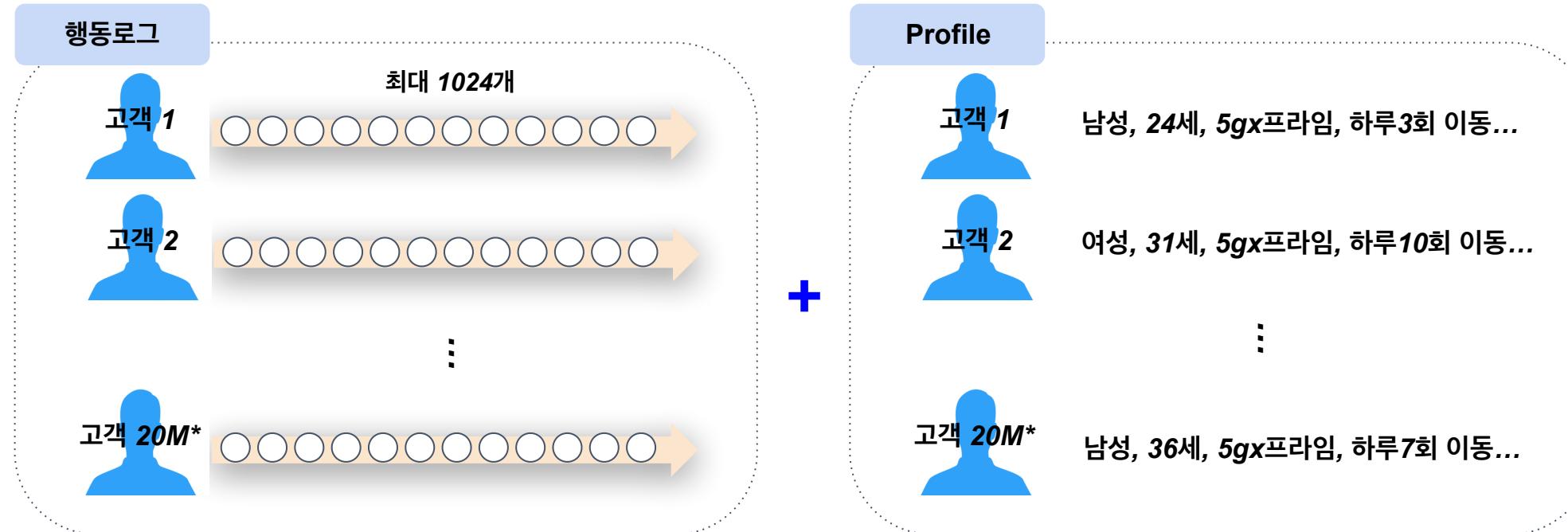


개발 현황 - 데이터

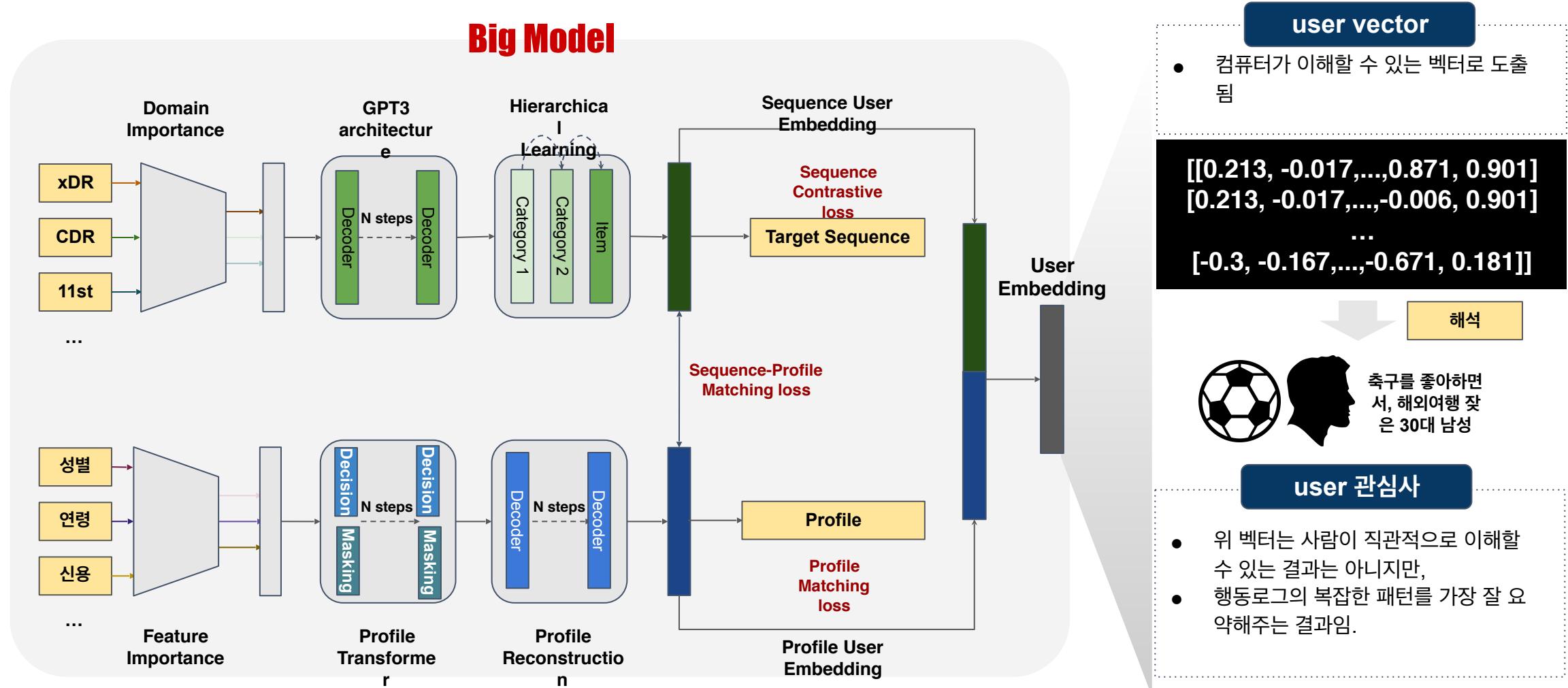
[데이터] 고객의 의도가 드러나는 행동로그를 중심으로,
성별, 연령 등 profile 정보와 함께 **통합 적재 완료** (약 300GB 이상)



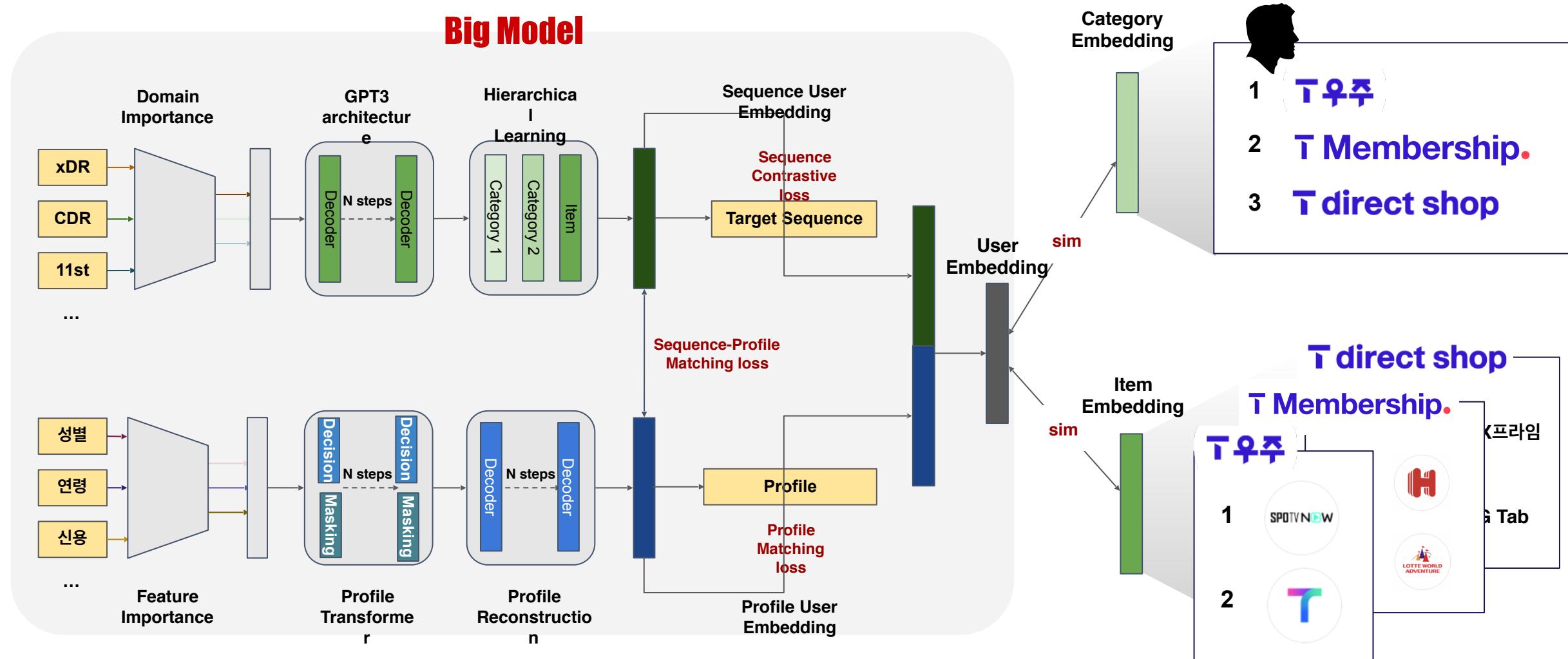
[데이터] 분석 동의자 1천6백만명의 정제된 고객 행동 로그 (평균길이: 420개) 와 더불어 profile 정보 (500가지 항목) 역시 통합 적재 완료



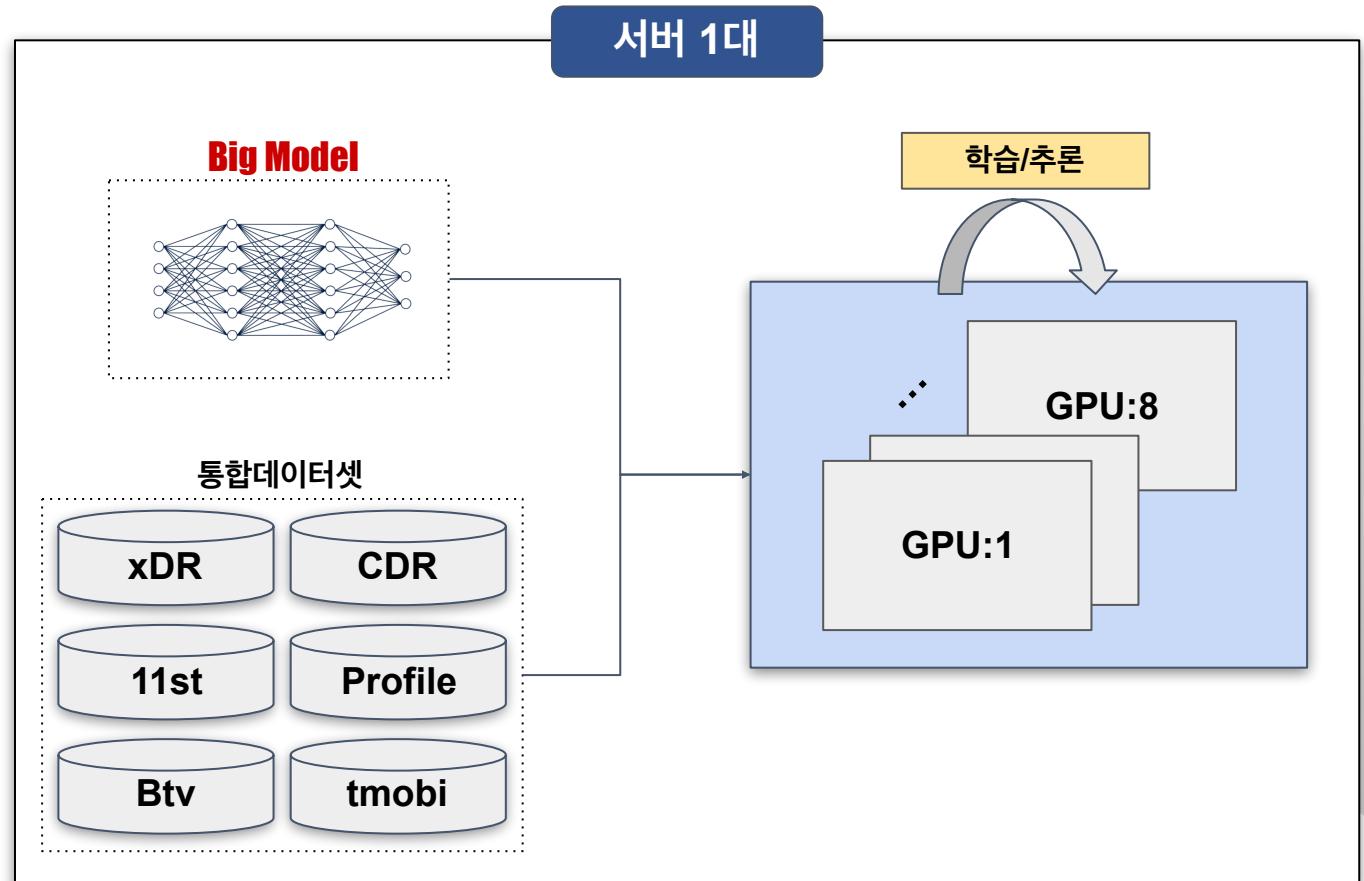
[모델] 다양한 영역의 추천/타겟팅에 접목할 수 있는 Big Model (Large Scale AI Model) 의 proto-type 개발 완료



[모델] 다양한 영역의 추천/타겟팅에 접목할 수 있는
Big Model (Large Scale AI Model) 의 proto-type 개발 완료



**V100 GPU (16GB) 8개의 서버 1 대로
300GB 분량 행동로그 + 4억개 parameter GPT 계열 모델 학습 (총 15일 소요)**



A.(에이디티)

A100 GPU (80GB) 1,040개

NAVER

A100 GPU (80GB) 1,120개

OpenAI

A100 GPU (80GB) 1만개

Meta

A100 GPU (80GB) 2,048개

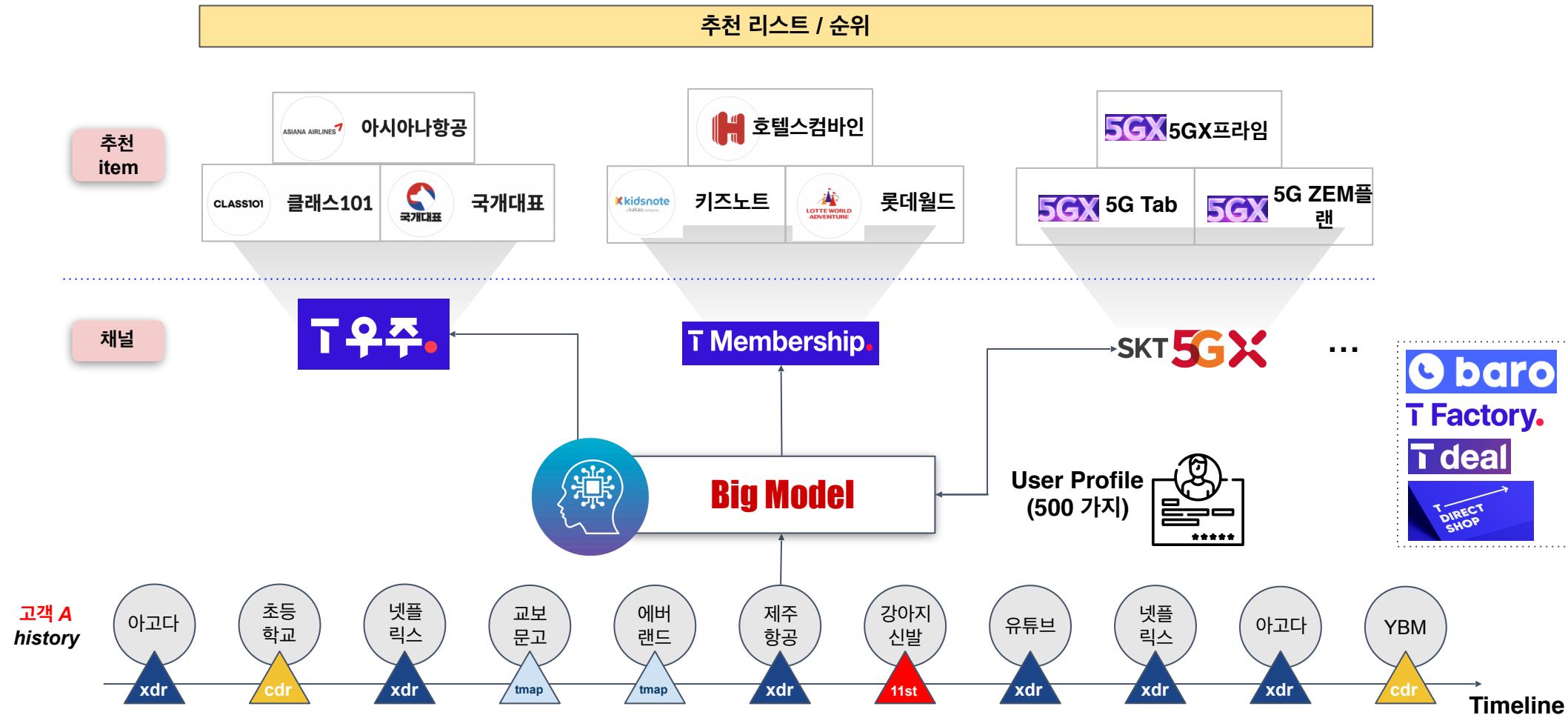
:



거대추천모델
프로젝트

V100 GPU (16GB) 8개

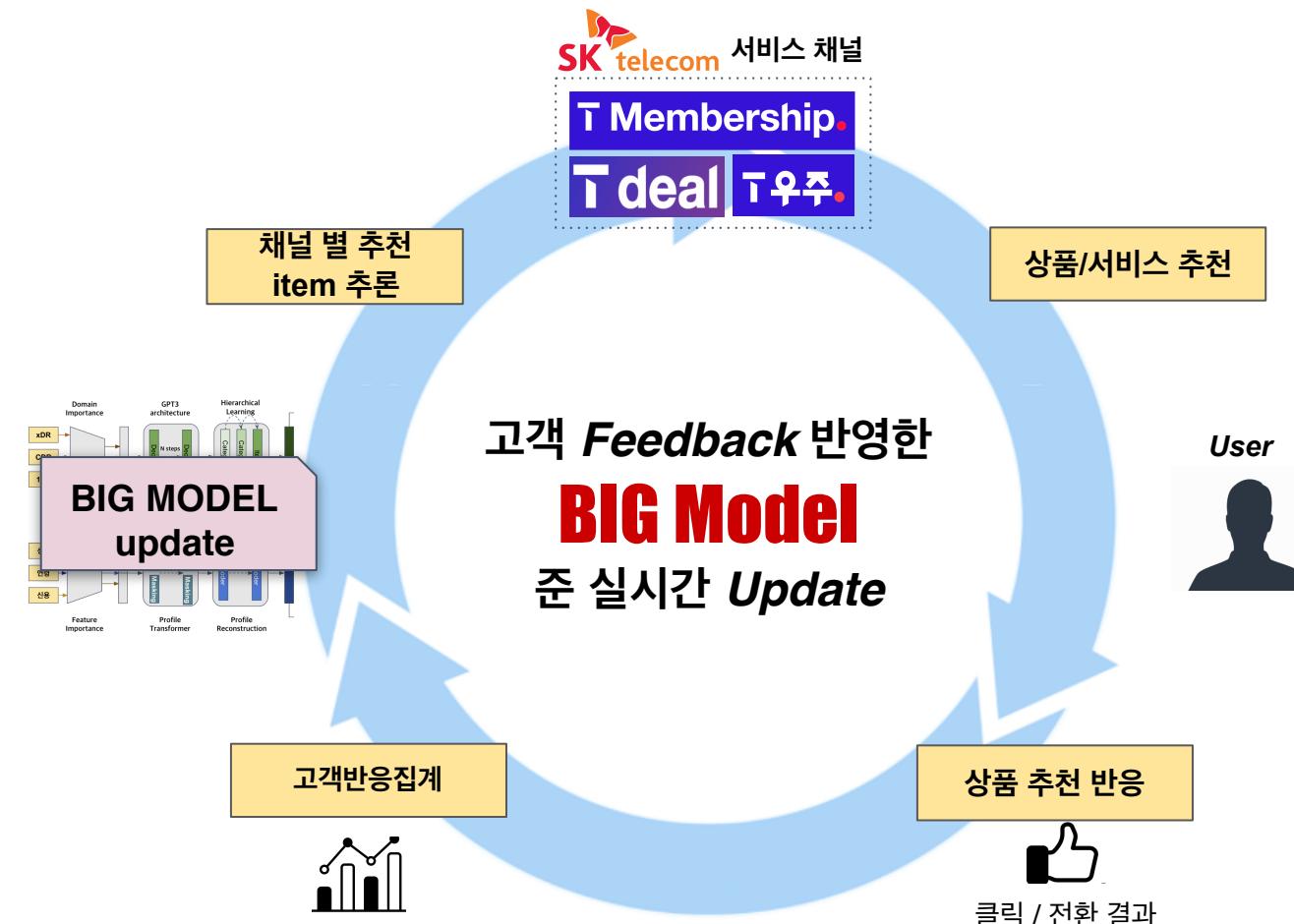
Big Model로 모든 채널 T우주, T-deal, MNO에 보다 정확한 추천을
통합적으로 진행 할 수 있음 (One-For-All)



개발 계획 ('23)- 고객 feedback기반 준실시간 모델 업데이트



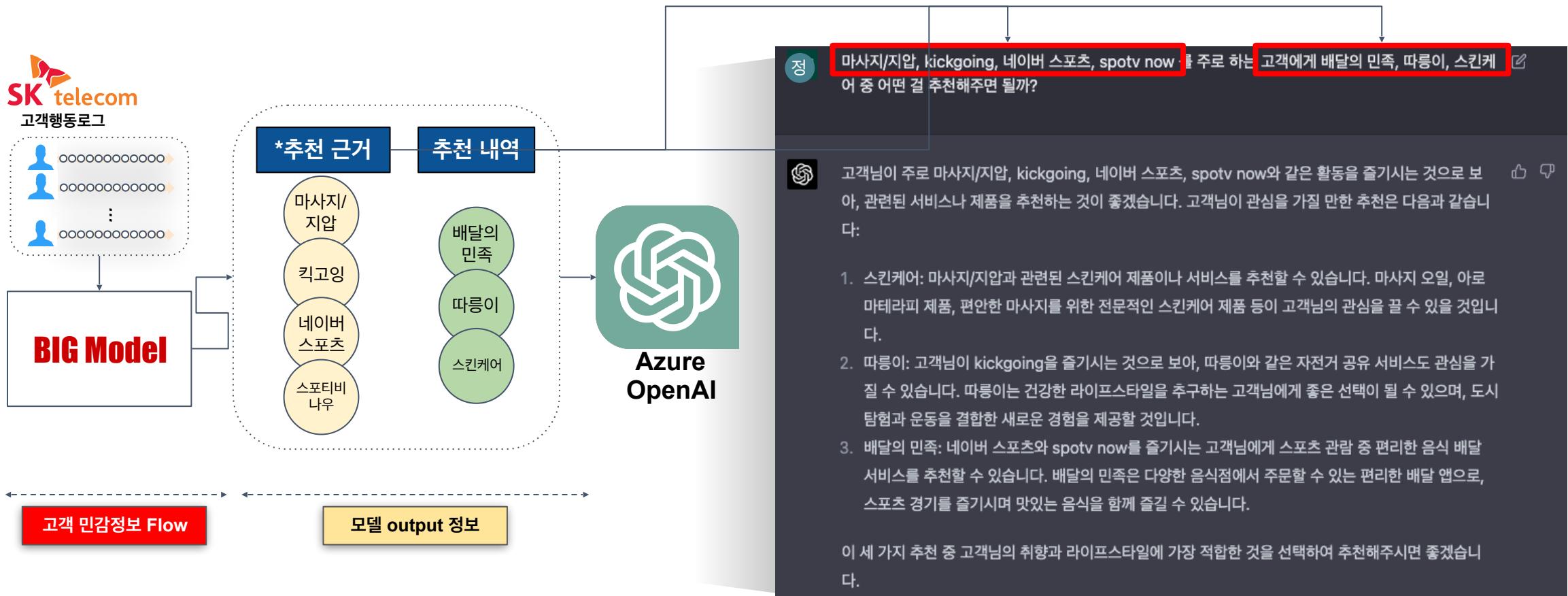
실시간 feedback 데이터를 모델에 반영/업데이트하는 모듈 개발하여
적시에 정확한 상품/아이템/타겟을 제공



특정 추천/타겟의 원인을 설명할 수 있는 Explainable module 개발



chatGPT를 연계하여 거대 추천 엔진의 추천 내역을 당사 상품과 매칭,
거대 추천 엔진의 추천 근거를 고객센터/유통망에 상품 추천을 위한 스크립트화 가능

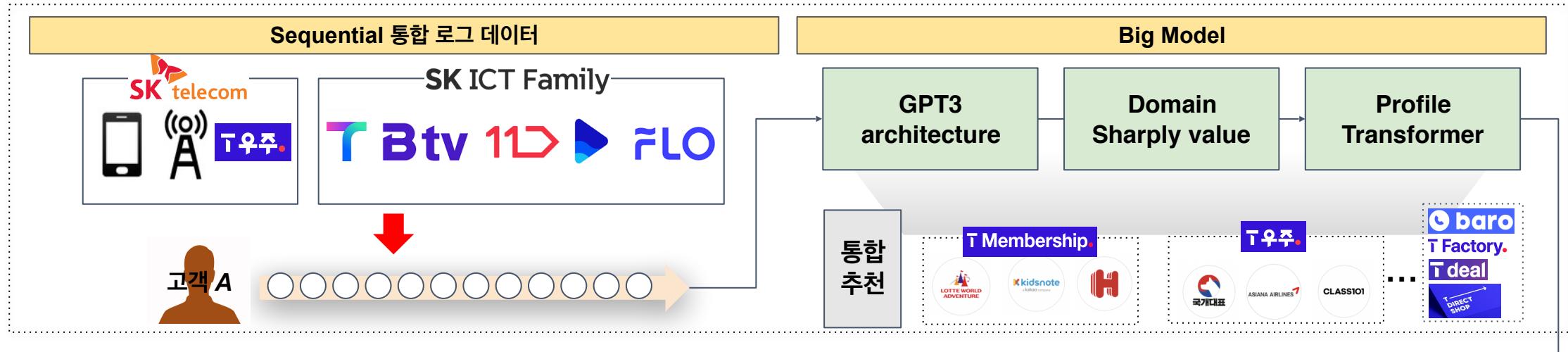


*추천 근거는 거대추천엔진의 output 중 하나로서 특정 추천의 결과에 가장 영향을 많이 준 행동 로그를 요약한 내용임

현재 개발 현황 summary



현재 BIG MODEL 개발을 위한 데이터 파이프라인 및 proto-type 모델은 개발 완료되었으며,
‘23.9월 이내 실제 마케팅에 적용 예정



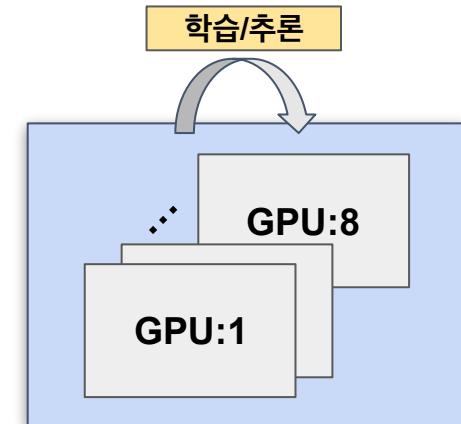
모든 domain의 로그들을 통합하고 시간 순서대로 정렬
- 고객의 관심사는 시간에 따라 변하므로 순차적 형태로 전처리

GPT기반 proto-type 모델 개발 (+데이터 domain 기여도 반영)
- 다음 로그 예측하는 GPT transformer decoder 구조를 backbone으로 하여, 데이터 domain (e.g., xdr) 기여도 기반으로 높은 성능의 모델 개발

ChatGPT 연계 검토

- 거대추천모델로부터 도출되는 추천 리스트 및 추천 근거를 프롬프트화하고 이를 토대로 추천 이유를 ChatGPT를 통해 스크립트화

[현재 서버 스펙]
*GPU : 8 V100 (16GB)
*CPU : 80 cores
*MEM : 1TB
*CUDA : 12.0



Contents

1. Why

2. Now

3. Future

Future ('24~)

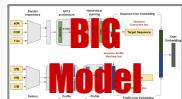
데이터 추가 학습 금융, 고해상도 PoI 등 및 모델 고도화를 통해 효율/활용성 높이고 전사 채널의 통합적 마케팅 관리 툴 개발 및 frontend에서의 ChatGPT 연계 강화

데이터/모델 고도화



금융, 고해상도 PoI 등 데이터 추가
→ 고객 행동 intent를 보다 정확히 파악

- * 마이데이터 사업으로 수집되는 금융/결제 데이터
- * 방문 PoI를 정확하게 알 수 있는 고해상도 PoI
- * ...



모델 구조 개선 및 scale-up → 보다 높은 정확도로 범용성 향상

- * 현재 4억개 parameters → 100억개 수준
- * 학습 속도 향상 (장비 추가 필요)
- * 준실시간성 모델 update 속도 향상

전사 마케팅 관리



고객 별 채널 최적화 및 피로도 관리→
보다 효율적이고 정확한 추천/타겟팅

- * 고객 피로도를 최소화하는 채널 배분
- * 고객 Feedback 데이터 인입/모델 업데이트 자동화를 통한 마케팅 효율 극대화



타겟팅/추천 효율 (매출 등) 극대화 하는 통합 관리 툴 개발

- * 전사 추천/타겟팅 모델을 일원화 할 수 있는 통합 관리 툴 개발
- * 매출 등 재무지표 극대화를 위한 강화학습 도입

ChatGPT 연계강화



거대추천엔진 추천결과와 ChatGPT
를 결합한 초기인화 메시징 제작

- * 모델 추천 결과 및 근거를 이용해 ChatGPT 미 세조정 학습으로 초기인화 메시징 도출



고객 페르소나 추론 및 유통망/고객센터 script 작성

- * 모델 추론 결과를 이용한 고객 페르소나 작성
- * 추천 결과의 해석의 script화

감사합니다!