데이터 없는 데이터 분석을 위하여
TrainDB: ML 모델 기반 근사 질의 처리 엔진

3세부

ETRI 한국전자통신연구원

TrainDB 개요

머신 러닝 기반으로 정확한 분석 결과에 준하는 근사 결과를 다양한 호스팅 환경에서 고속 제공하는 탐사 데이터 분석 지원 DBMS 근사 질의 처리 엔진 기술

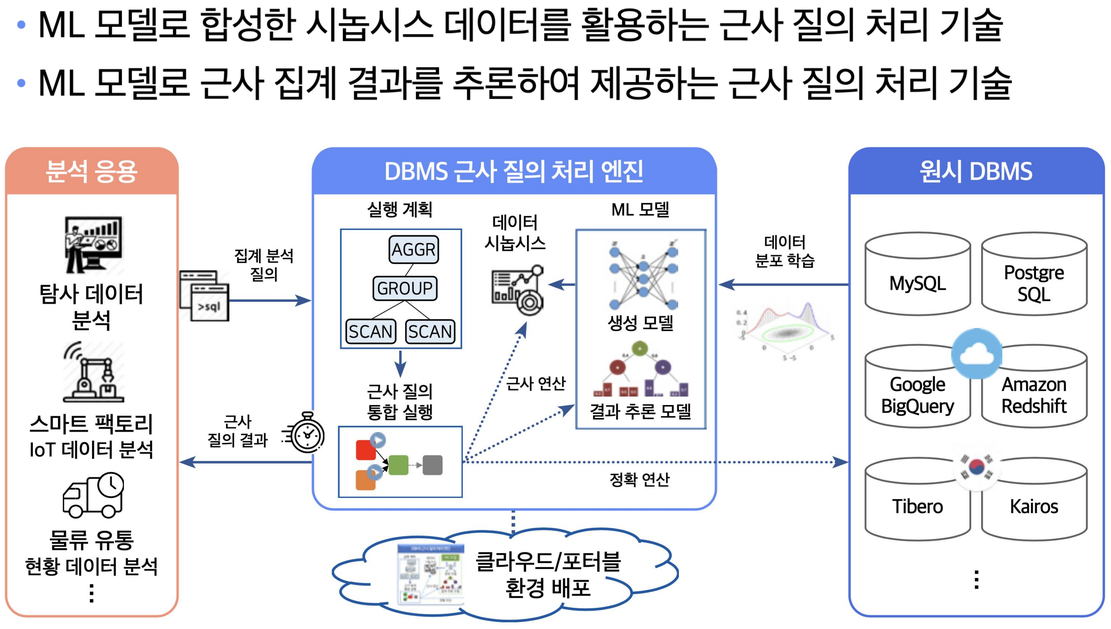
TrainDB 특징

- ML 모델을 학습한 이후에는 원시 데이터에 접근하지 않고 근사 질의 결과 제공
- Kubeflow 기반 클라우드 머신 러닝 서비스와 연동하여 원격 ML 모델 학습/실행 가능
- 다양한 DBMS를 원시 DBMS로 연동하여 데이터 분석 질의 수행

핵심 기술

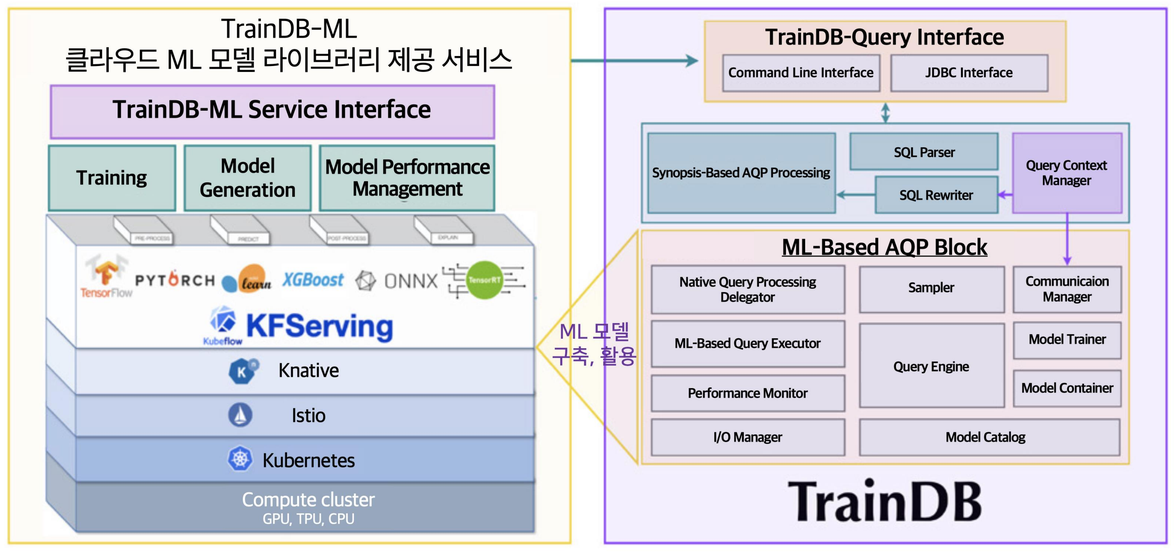
TrainDB: ML 모델 기반 근사 질의 처리 엔진

- SQL 기반 근사 질의 언어 확장
- ML 모델로 합성한 시놉시스 데이터를 활용하는 근사 질의 처리 기술
- ML 모델로 근사 집계 결과를 추론하여 제공하는 근사 질의 처리 기술



클라우드 ML 모델 라이브러리 제공 서비스

- 원격 GPU 서버에서 ML 모델을 학습/제공하기 위한 프레임워크
- 클라우드에서 작동 가능한 Kubeflow 기반 ML 모델 등록/학습/실행 기능 지원



시연 ML 모델을 이용한 근사 질의 실행

사전 준비 단계

CREATE MODELTYPE → TRAIN MODEL → CREATE SYNOPSIS

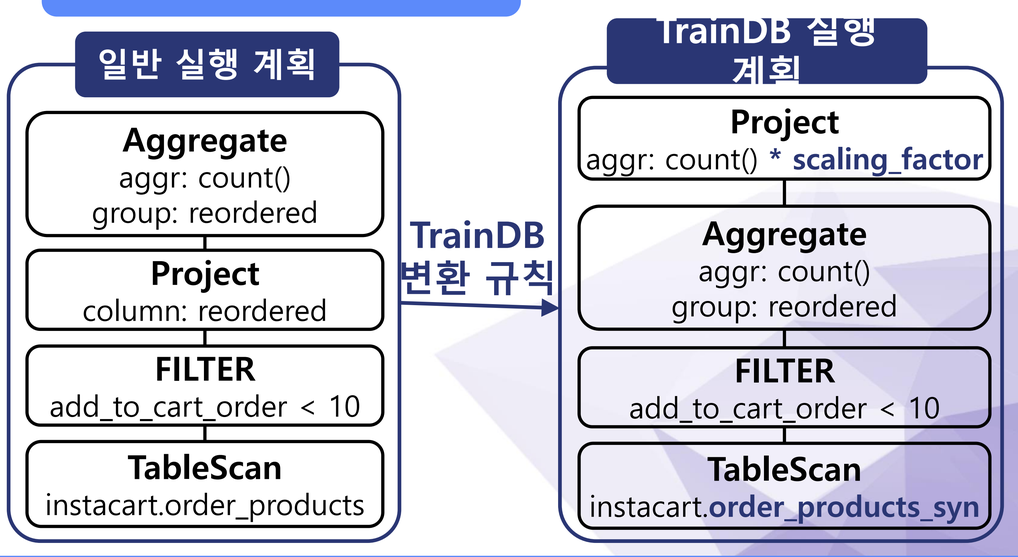
ML 모델 유형 등록 | 테이블 데이터로 ML 모델 학습 | ML 모델로 시놉시스 생성

실행 단계

SQL 질의문

```
SELECT APPROXIMATE reordered, count(*)
FROM instacart.order_products
WHERE add_to_cart_order < 12
GROUP BY reordered;
```

TrainDB Planner



정확 질의와 근사 질의 비교 사례 – 실행 시간과 결과 비교

```
+-----------+----------+
| reordered |  EXPR$1  |
+-----------+----------+
| 1         | 15137795 |
| 0         | 9231874  |
+-----------+----------+
2 rows selected (13.742 seconds)
```

< 정확 질의 >

```
+-----------+----------+
| reordered |  EXPR$1  |
+-----------+----------+
| 1         | 15281242 |
| 0         | 10328628 |
+-----------+----------+
2 rows selected (0.302 seconds)
```

< 근사 질의: 1% 크기의 시놉시스 >