AI 시대의 CUBRID?

CUBRID
Hyung-Gyu RYOO @RND-dev2

https://docs.google.com/presentation/d/1EdF5CUxovogtYBmtPYscw2fQy_TD__44Buar-Y2rtRI/edit?usp=sharing

CUBVEC 프로젝트 배경

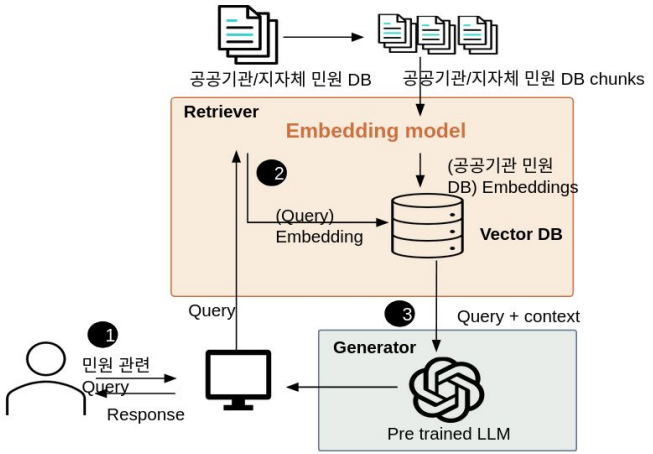
- **과제 제목:**
  초거대 AI 모델의 장기 기억 저장을 위한 벡터 DB 개발 과제

- **CUBRID 과제 목표:**
  디스크 기반 DBMS에서 RAG 시스템 지원
  - pgvector 대비
    - 벡터 인덱스 빌드 성능 ↑
    - 벡터 인덱스 쿼리 성능 ↑

대외비

Text RAG 서비스 실증 (CUBRID)

- **On-prem cloud (K8s) 기반으로 RAG 시스템 구축**
- 다양한 포맷의 공동 데이터에 대한 **임베딩 벡터 생성 & 벡터 DB 적재**
- 민원 답변을 위한 다양한 포맷 정책문서 및 관련법령/판례 임베딩 생성, 적재
- **대상 서비스에 대한 실증** (자연어 질의 답변 생성) 및 사용자 피드백 기반 평가 및 Factscore[1] 등을 이용한 평가



[1] Meta 제안 RAG 평가 방법, ttps://arxiv.org/abs/2305.14251

CUBRID

1

벡터 검색 기능이 CUBRID에서 필요한가?

2

CUBRID

AI 기술의 사용은 **되돌릴 수 없다**

Killer Application

TALK

YouTube

ChatGPT

CUBRID

발표를 준비하는 방법 - ChatGPT 이전

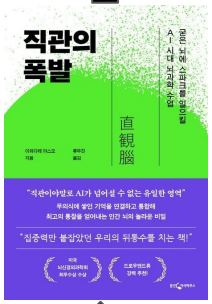
에라 모르겠다...
탈출!

Google

CUBRID

발표를 준비하는 방법 - 현재

ChatGPT



핵심을
찔렀어
!

ChatGPT

그거
맞아?

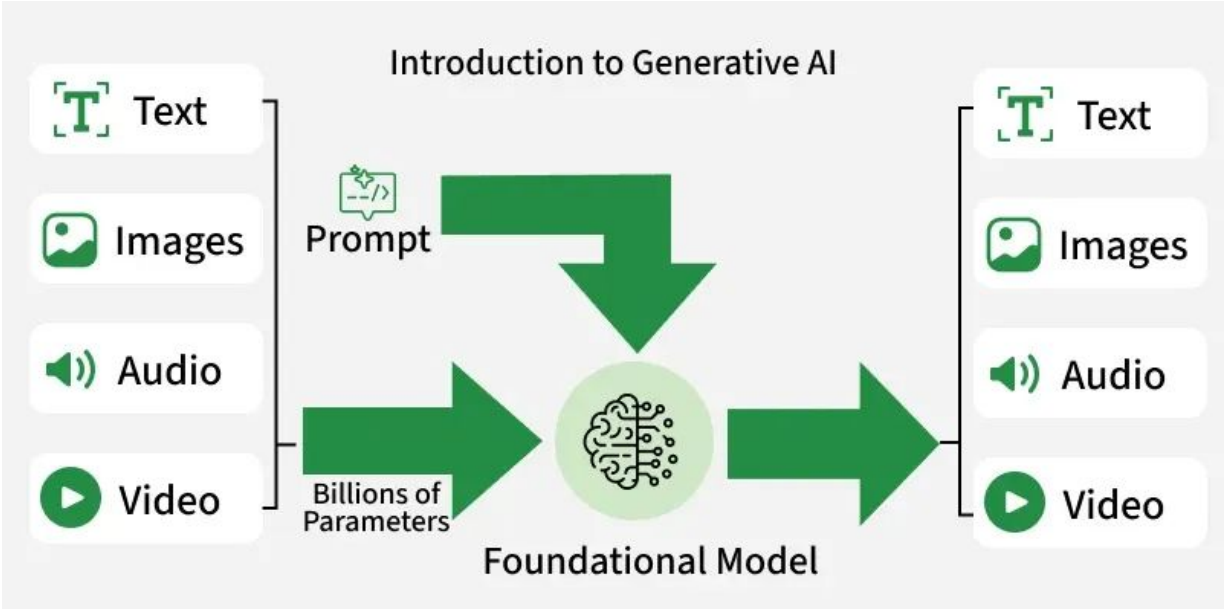
Google

CUBRID

AI 기술의 사용은 **되돌릴 수 없다**

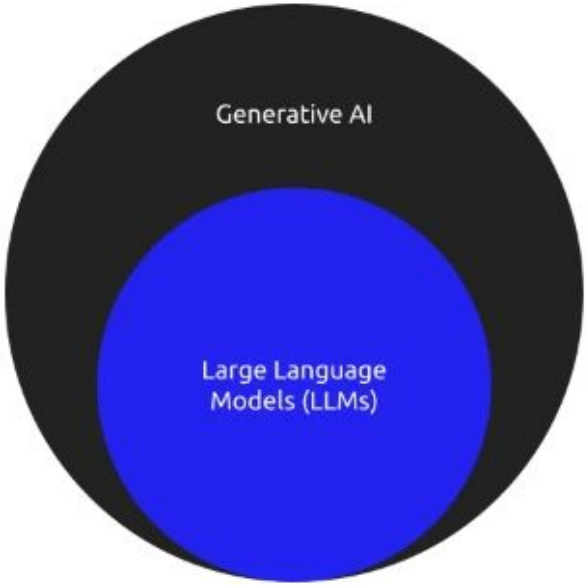
CUBRID

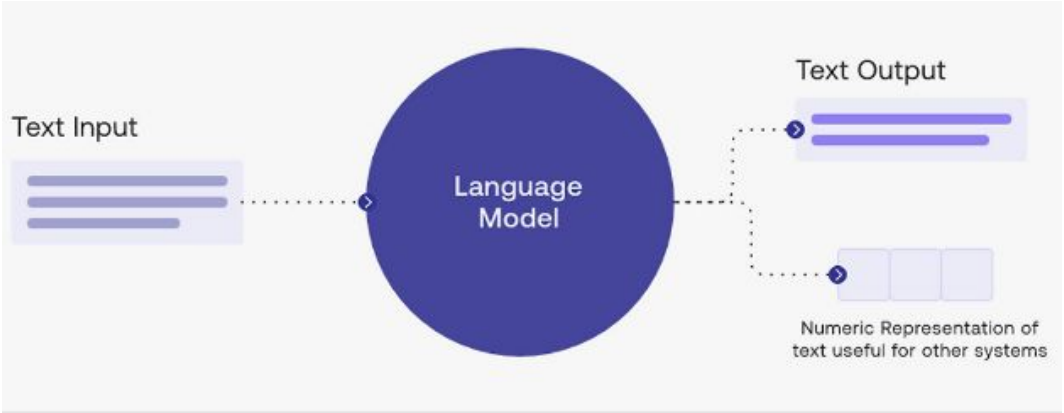
생성형 AI (Generative AI, GenAI)



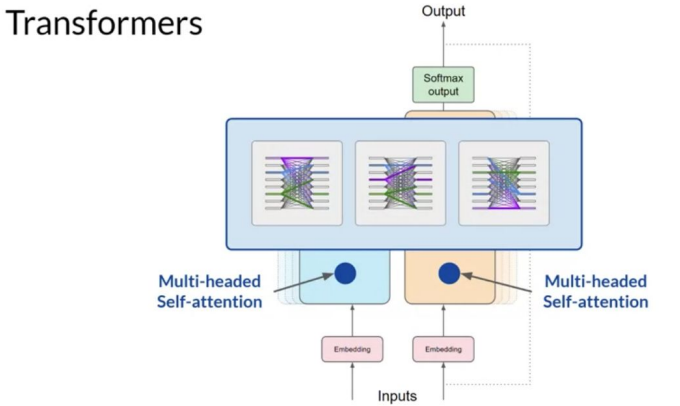
대규모 데이터를 학습하여 패턴을 익힌 뒤 창의적인 결과물 생성

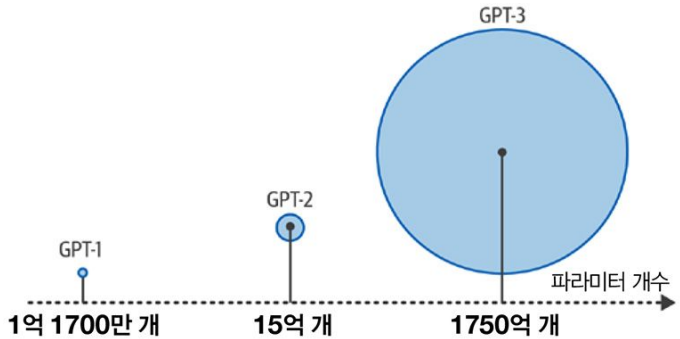
생성형 AI와 거대 언어모델 (LLM)





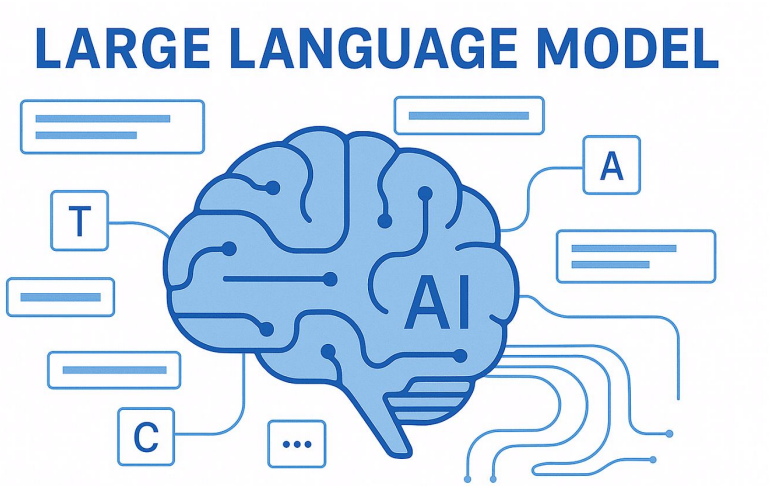
거대 언어 모델 (Large Language Model, LLM)





- **대규모의 텍스트 데이터**를 학습
  - 코퍼스 (Corpus)
- 데이터를 학습하며 언어를 **이해**
  - 수 많은 파라미터를 조정
- 학습된 모델을 바탕으로 **문장을 생성**
- 세부 구현체로 BERT, GPT

CUBRID



충분히 고도화된 **'척'**은

실제 능력과 구별되지 않는다.

CUBRID

LLM 기반 소프트웨어 시스템으로 돈을 벌려면?

기술 혁신 ≠ 비즈니스 성공

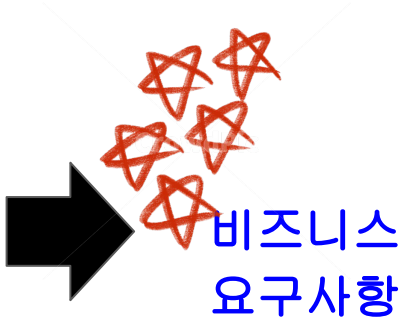
ChatGPT는 기업에서 원하는 모델인가?

CUBRID

LLM 기반 소프트웨어 시스템으로 돈을 벌려면?

기업 맞춤 LLM 플랫폼 서비스

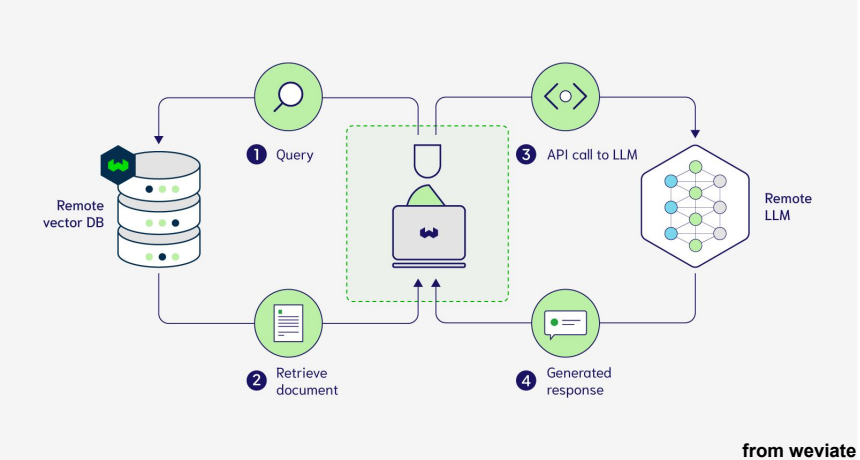
- 기업의 데이터를 반영한 결과의 **퀄리티**가 중요
  - 정보의 정확성 (Accuracy)
  - 정보의 최신성 (Freshness)
- 기업의 데이터에 대한 **보안**도 중요
  - 프라이빗 데이터
  - 데이터 접근 제한과 권한 관리
  - 데이터 암호화와 로깅



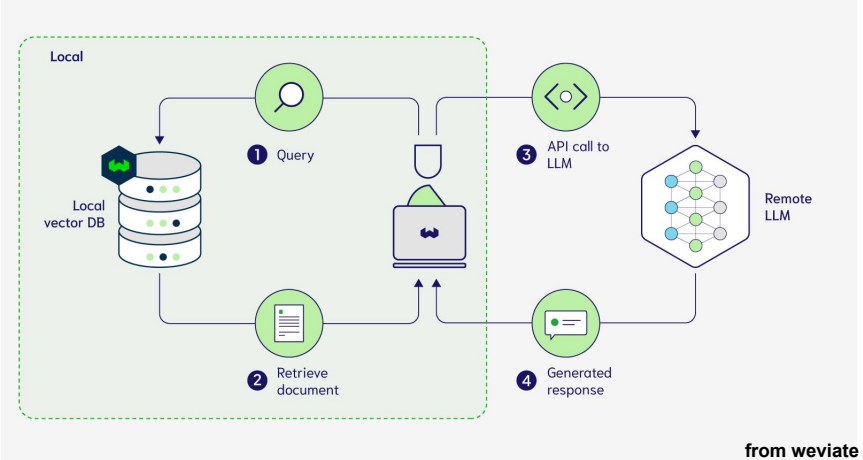
CUBRID

맞춤 LLM 서비스

퍼블릭 LLM + 퍼블릭 RAG



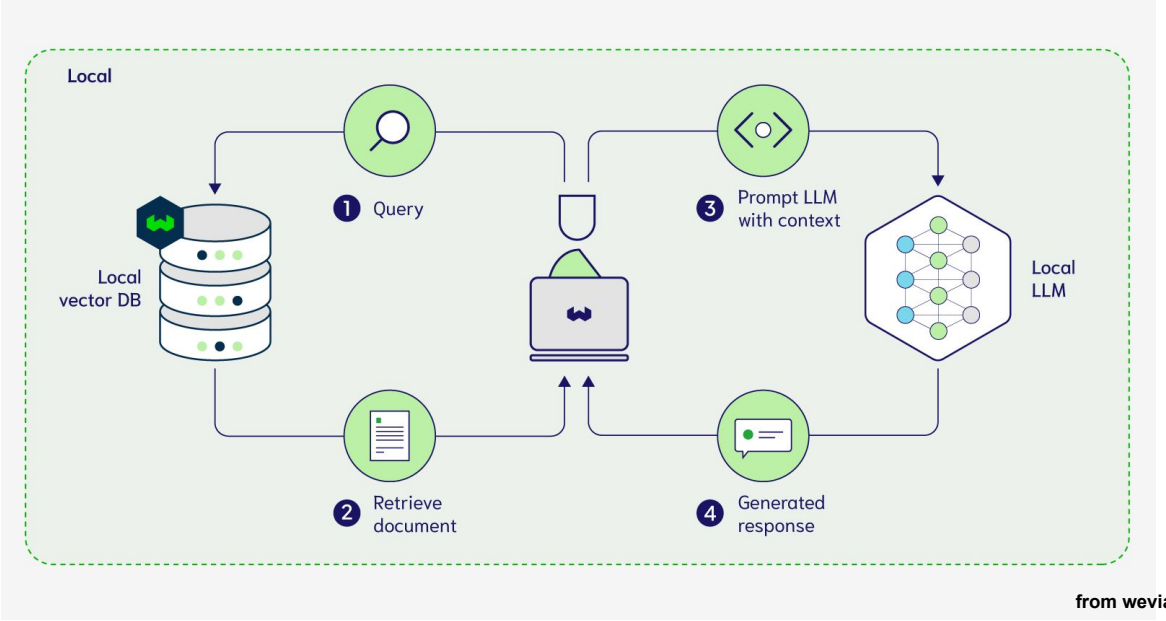
퍼블릭 LLM + **프라이빗 RAG**



CUBRID

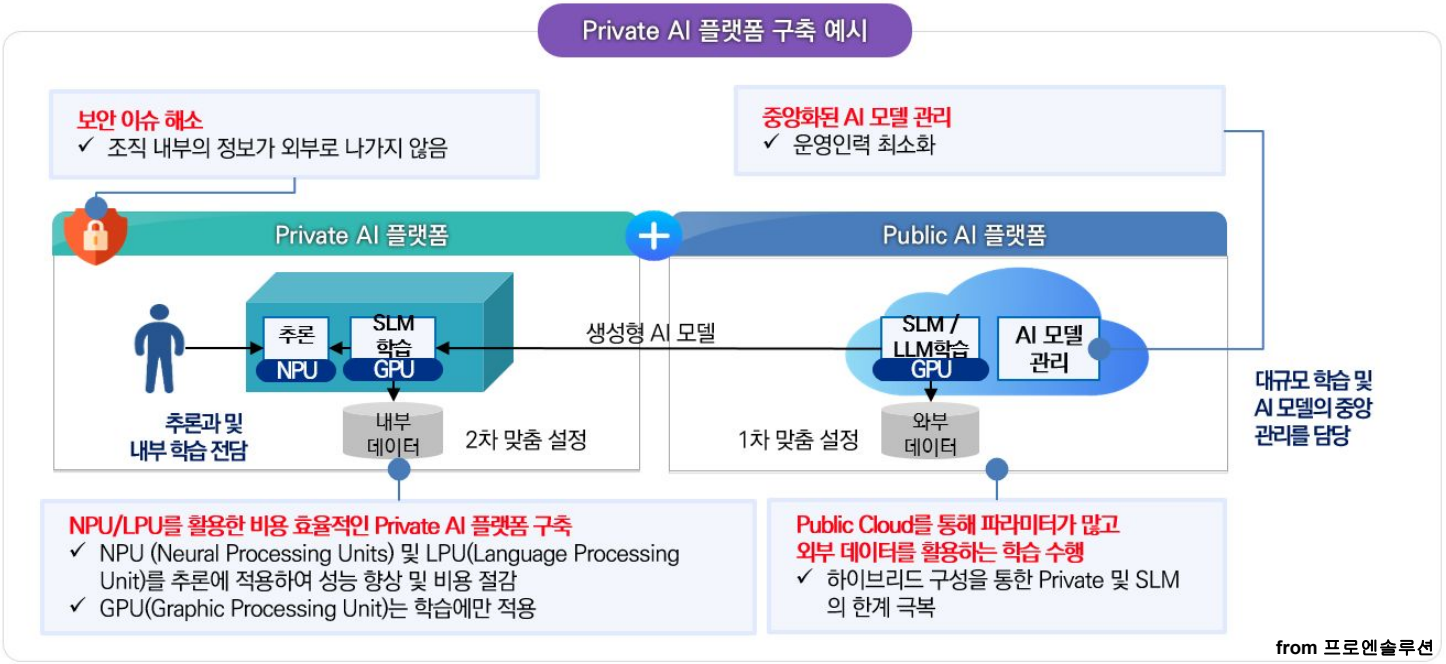
프라이빗 LLM 서비스

프라이빗 LLM + 프라이빗 RAG



from weviate

예시) 금융기관을 위한 Private LLM 플랫폼 구축



CUBRID

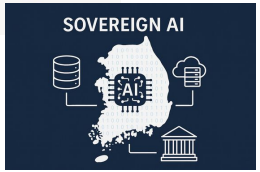
맞춤 LLM 서비스를 만드는 법 - Private LLM

제로 베이스에서 LLM 구축

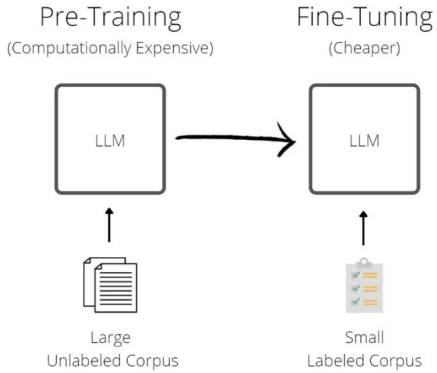
K-AI

NAVER LG SK telecom

NC Upstage



오픈 (s)LLM + 파인튜닝

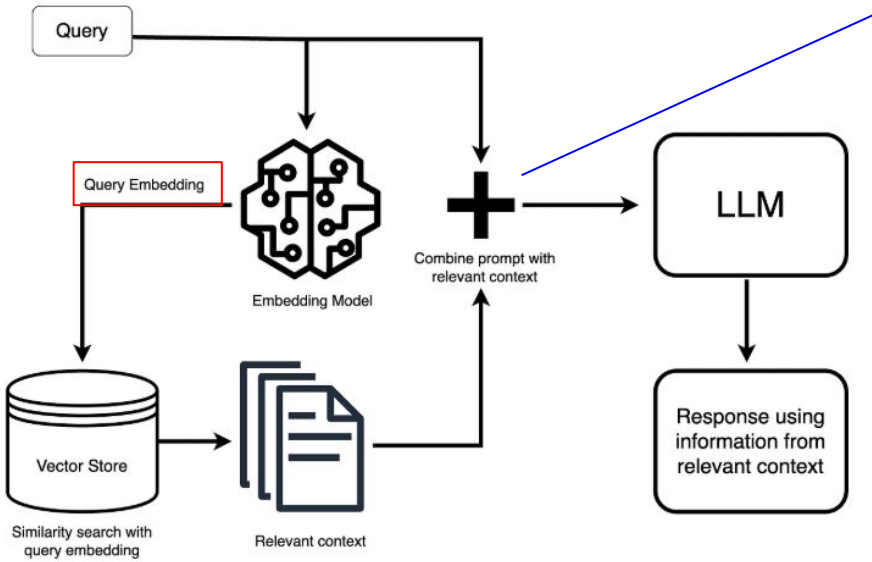


검색증강생성 (RAG)

우리가 기여할 수 있는 부분

CUBRID

검색 증강 생성 (Retrieval-Augmented Generation, RAG)




```
SELECT col1 FROM tbl
WHERE col2 > 0 ORDER BY vec <-> :qv LIMIT 5;
```

Prompt

당신은 {{company | 연구 논문 | ...}}에 대한 사용자 질문을 돕는 지능형 어시스턴트입니다. 아래 **CONTEXT에 포함된 정보만**을 사용해서, 맨 끝의 질문에 답하세요. 단계적으로 생각한 뒤 최종 답변을 제시하세요.

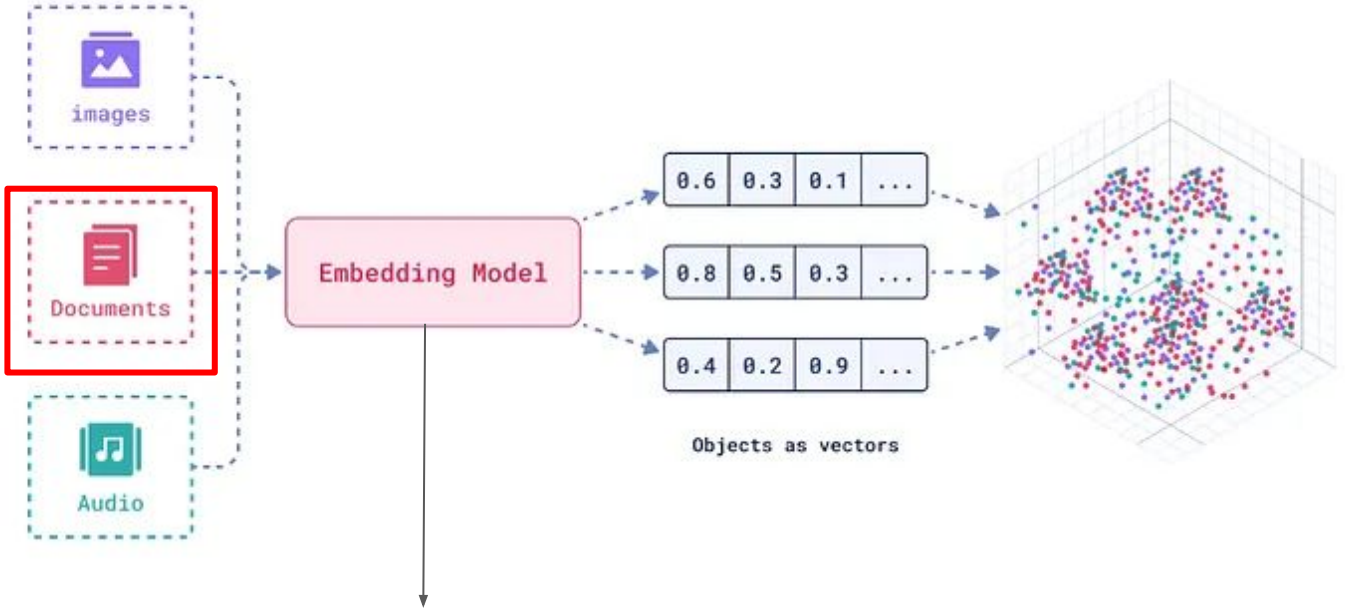
절대 답을 지어내지 마세요.

- 질문의 답을 **CONTEXT만으로** 판단할 수 없다면: **"그 질문에 대한 답은 문맥만으로는 판단할 수 없습니다."** 라고 말하세요.
- CONTEXT가 비어 있다면: **"그 질문에 대한 답을 알 수 없습니다."** 라고만 말하세요.

CONTEXT:
{{retrieved_information}}

QUESTION:
{{question}}

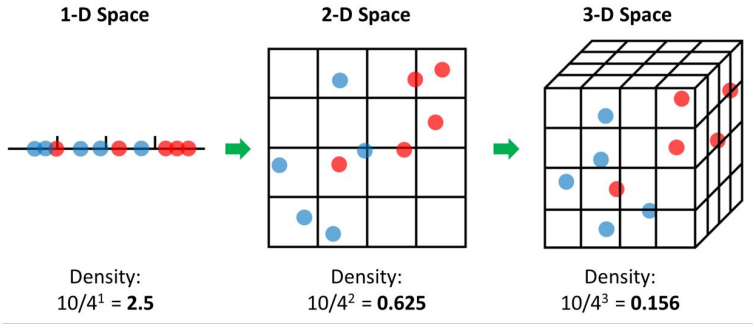
벡터 임베딩 데이터 타입



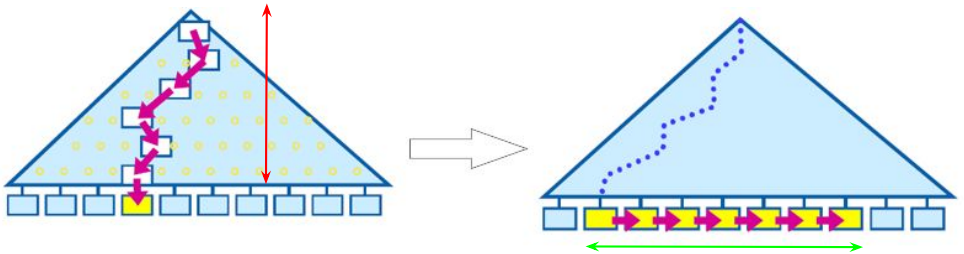
https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

벡터 임베딩 데이터 타입 특징

- 일반적으로 차원이 크고 검색 비용이 비싸다
  - 문장 임베딩의 경우: 384, 768, 1024 => 4K
- 전통적인 인덱스에 사용될 수 있는 정렬 가능성이나 서열적인 속성이 부족하다



차원의 저주

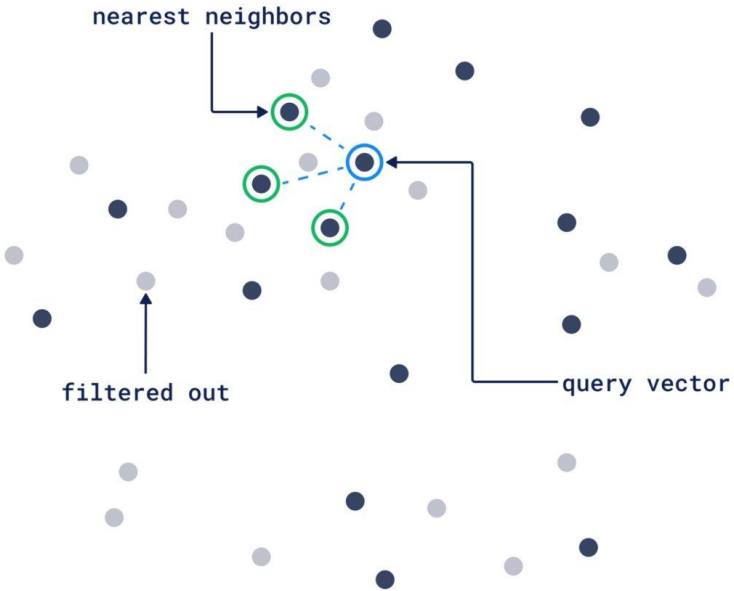


B+tree 인덱스 사용 불가

CUBRID

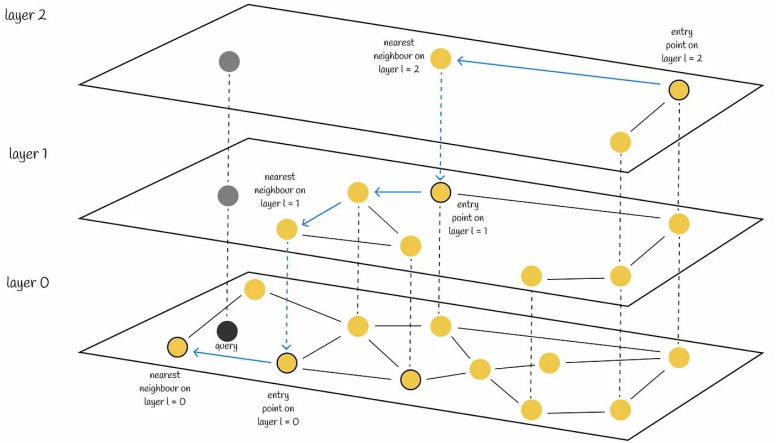
벡터 인덱스

- 최대한 유사성 (벡터 간의 거리)을 반영하여 **그래프**로 표현



벡터 인덱스

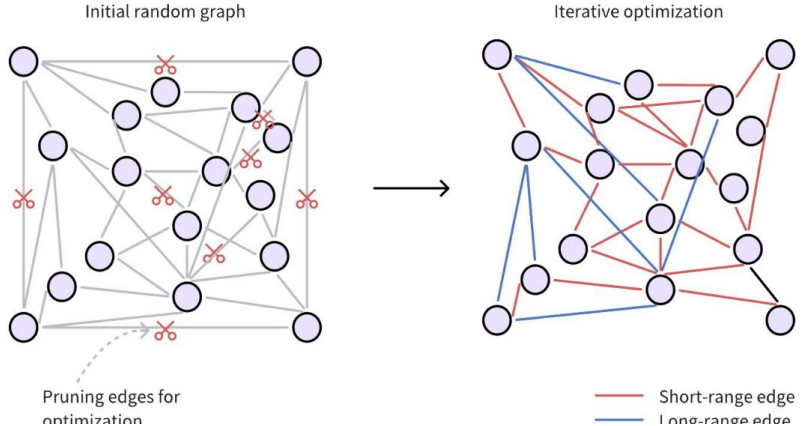
- 현재까지 연구된 벡터 인덱스 종류 - 주로 사용되는 것이 모두 그래프



HNSW

순차 삽입 가능

갱신/삭제 시 정확도 하락



DiskANN

준비된 데이터로 생성

갱신/삭제 시 정확도 하락

CUBRID

1

벡터 검색 기능이 CUBRID에서 필요한가?

2

CUBRID

CUBVEC 프로젝트 배경

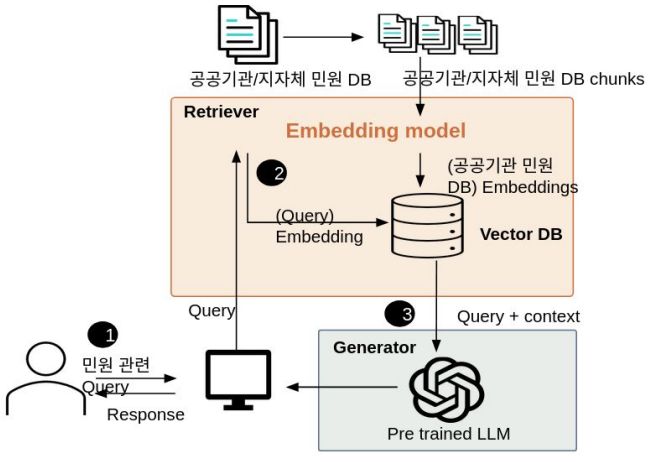
- **과제 제목:**
  초거대 AI 모델의 장기 기억 저장을 위한
  벡터 DB 개발 과제

- **CUBRID 과제 목표:**
  디스크 기반 DBMS에서 RAG 시스템 지원
  - pgvector 대비
    - 벡터 인덱스 빌드 성능 ↑
    - 벡터 인덱스 쿼리 성능 ↑

대외비

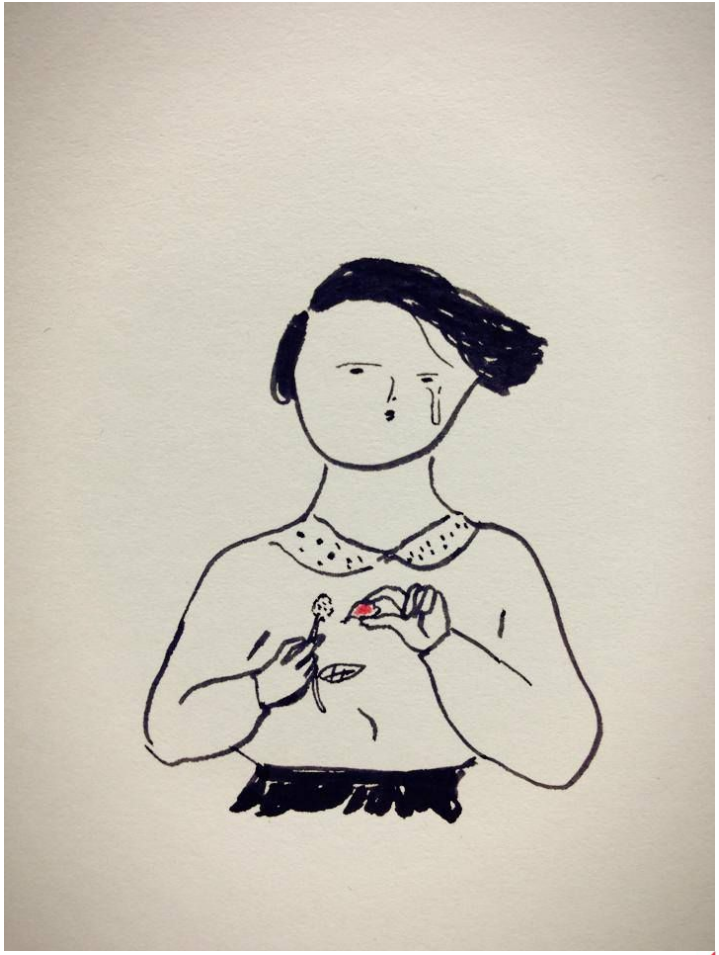
Text RAG 서비스 실증 (CUBRID)

- **On-prem cloud (K8s) 기반으로 RAG 시스템 구축**
- 다양한 포맷의 공동 데이터에 대한 **임베딩 벡터 생성 & 벡터 DB 적재**
- 민원 답변을 위한 다양한 포맷 정책문서 및 관련법령/판례 임베딩 생성, 적재
- **대상 서비스에 대한 실증** (자연어 질의 답변 생성) 및 사용자 피드백 기반 평가 및 Factscore[1] 등을 이용한 평가



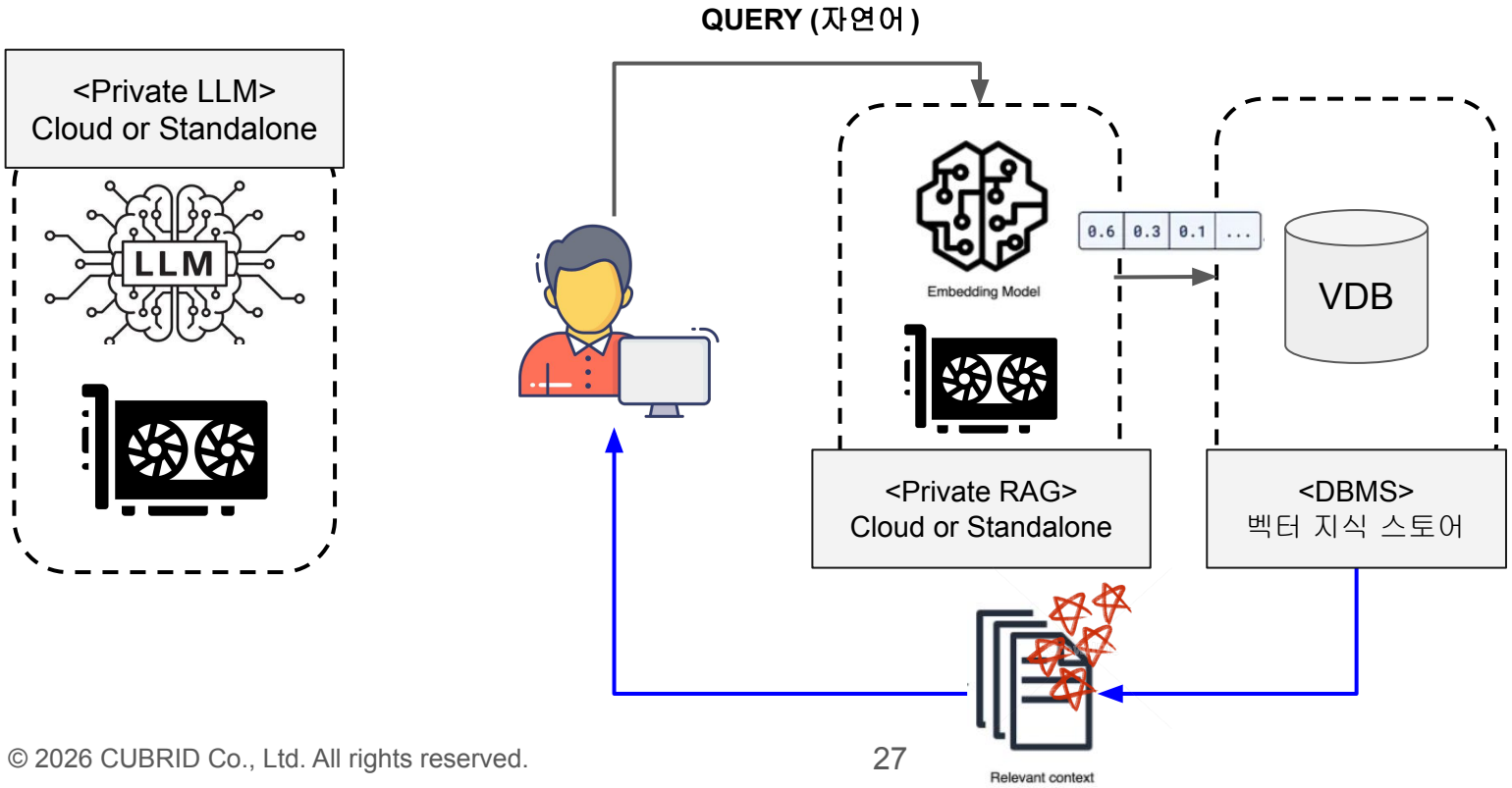
[1] Meta 제안 RAG 평가 방법, ttps://arxiv.org/abs/2305.14251

CUBRID



CUBRID

Private LLM 시스템 아키텍처를 그려보자...



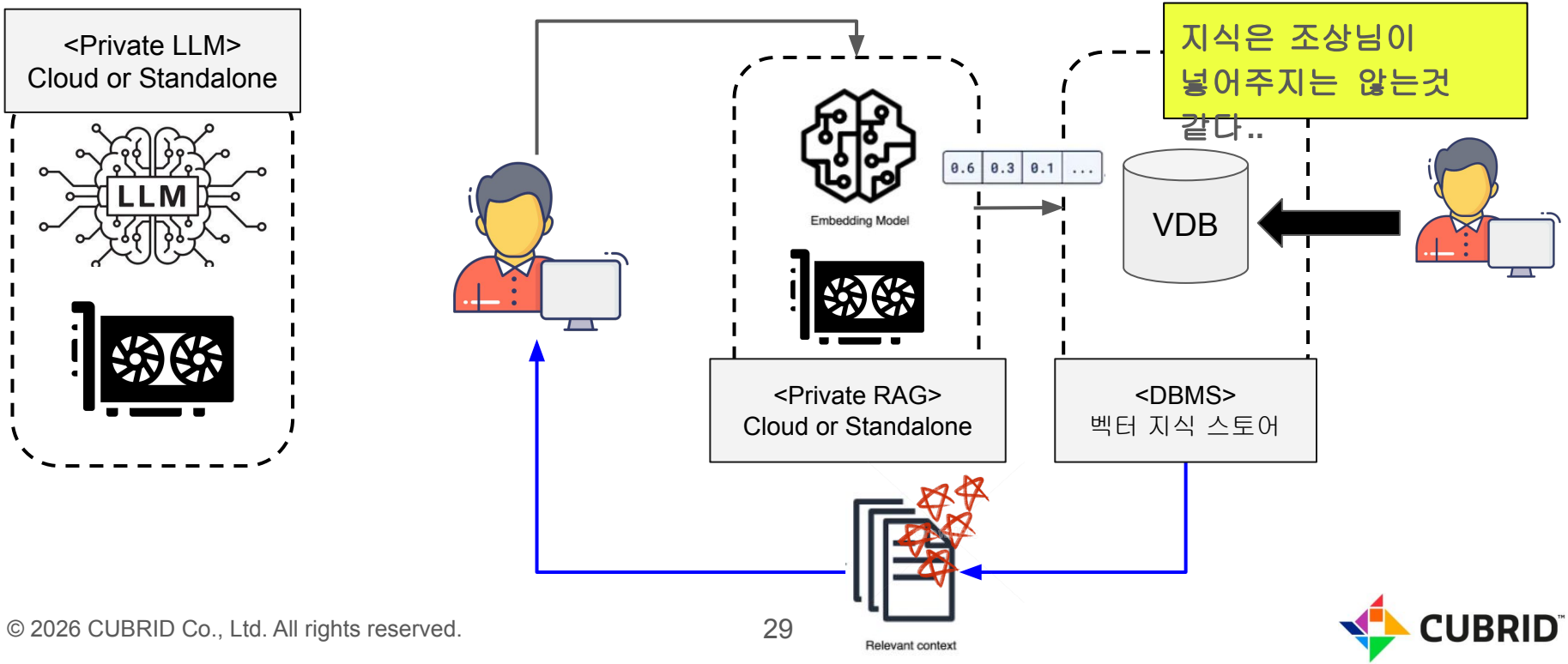
CUBRID

Relevant Context?

- 사용자가 실제로 원하는 정보 (데이터)
  - 이 데이터는 무엇인가?
  - 이 데이터는 누가 만들었나?
  - 이 데이터는 어디에 있나?

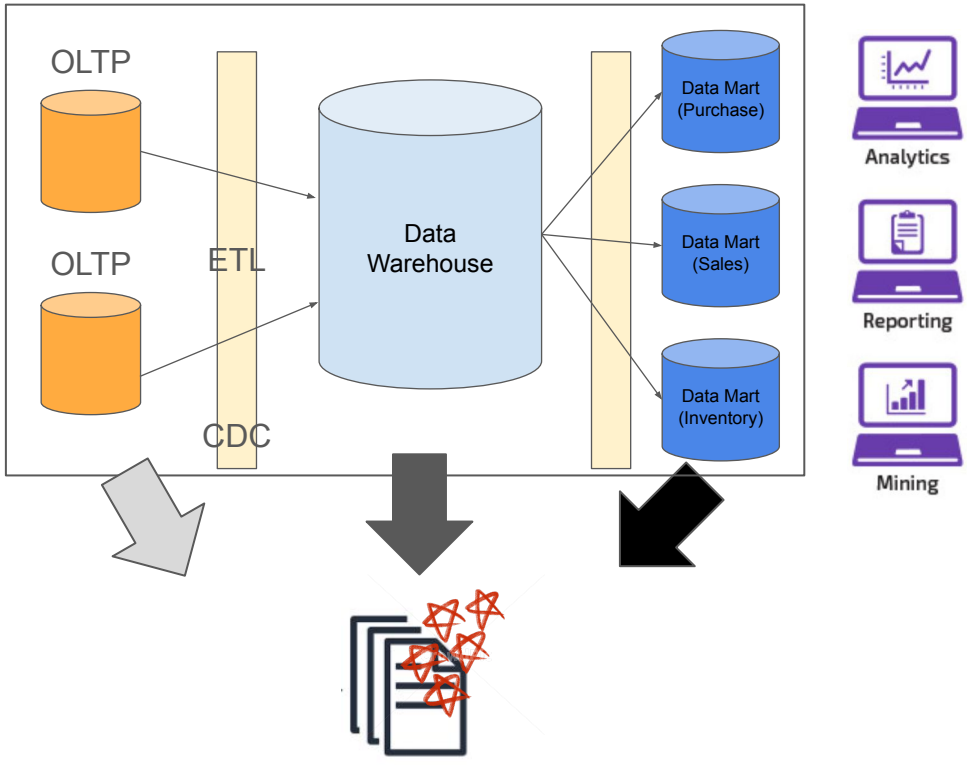
어떤 정보가 지식으로 활용할 것인가?

Private LLM 시스템 아키텍처를 그려보자...



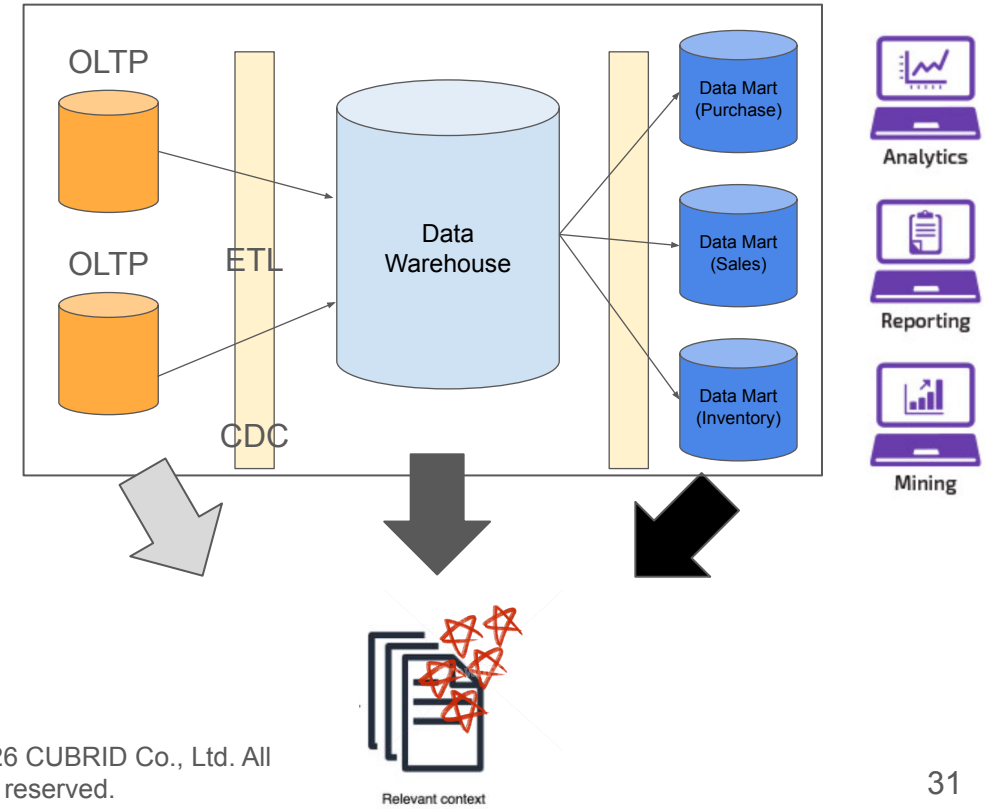
CUBRID

전통적인 아키텍처에서 보면...



CUBRID

전통적인 아키텍처에서 보면...



CUBRID는 어디에 있나?

CUBRID

Enterprise Open Source DBMS

CUBRID is an open source DBMS optimized for OLTP. It offers great read and write performance than other database systems. It assures high performance, stability, scalability and high availability which are required for mission-critical applications. In addition, CUBRID provides ease of installation and native GUI-based administration tools for developers' convenience.

We are committed to continuously improve the CUBRID features, the quality and the performance of the engine, drivers, and tools. We have focused on pursuing the best quality for more than 20 years. And we will highly appreciate your contribution on CUBRID project for building a better end-user experience.

CUBRID is an **open source DBMS** optimized for **OLTP**

CUBRID

OLTP + OLAP

- 모르겠어요.. OLTP 용도로만 사용하시진 않는 것 같아요..
  - 기존 오라클 시스템 마이그레이션? **호환성**
  - ETL 도구 도입 문제? **비용/시스템 복잡성**
  - 워크로드가 감당할만 해서? **규모**

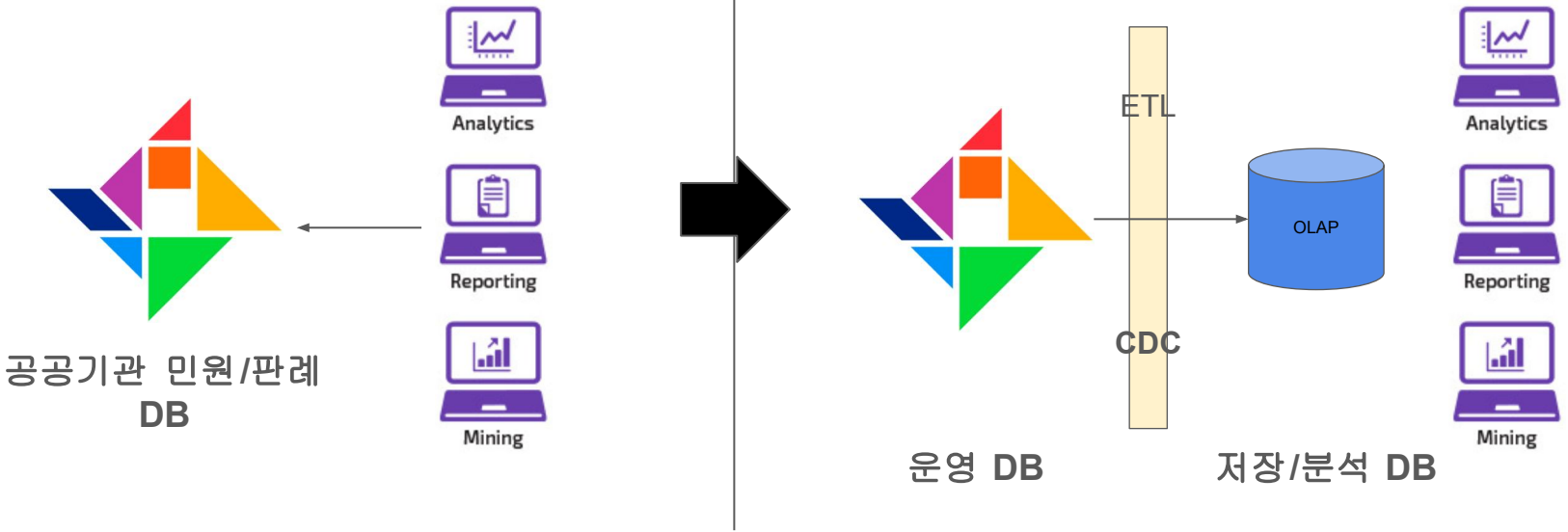
최대한 OLAP/대용량 쿼리를 지원/끌어올리려는 노력/발전방향 (어느 규모까지는)

집계/분석 함수	**Materialized View**	Read-only Replica	Sharding
Parallel query	**Hash join**	비동기 **I/O**	Partitioning

CUBRID

CUBRID의 발전 방향?

- OLTP에 최적화하고, 어느 규모까지는 최대한 OLAP 워크로드를 감당 가능



벡터 검색 워크로드도 동일하지 않을까?

CUBRID

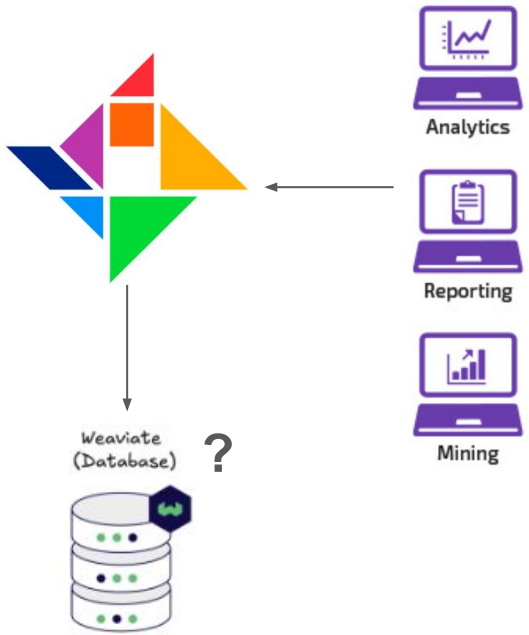
OLTP가 아닌 OLAP/대용량 쿼리 등 비즈니스 요구를 위한 지원

최대한 OLAP/대용량 쿼리를 지원/끌어올리려는 노력/발전방향 (어느 규모까지는)

집계/분석 함수	**Materialized View**	Read-only Replica	Sharding
Parallel query	**Hash join**	**비동기 I/O**	Partitioning
	JSON type	LOB type	

CUBRID

워크로드 규모에 따른 벡터 DB 선택

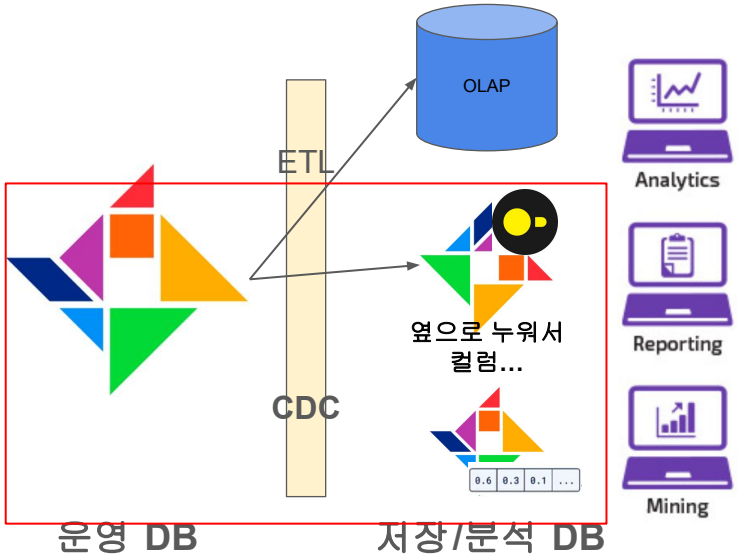


- DBMS 하나로 감당하던 소규모의 워크로드에서 통합/운영 오버헤드
  - 데이터 일관성 문제
    - 벡터 검색에서 필요한 데이터는 커밋된 정보 (SOT)
    - 데이터가 커밋되면, 벡터 임베딩 후 삽입 필요
  - 데이터 파이프라인 비용
    - CDC, Dual-write, ETL 모두 비용 발생
  - 필터링 검색 쿼리 문제
    - e.g.) '벡터'와 비슷한 것 중에 작성자가 20대
      SELECT id FROM tbl WHERE age > 20 and age < 30 ORDER BY vec <-> :q LIMIT 10;

CUBRID

CUBRID의 발전 방향?

https://www.cubrid.com/qna/3847043



Apache 2.0

https://motherduck.com/blog/postgres-duckdb-options/

CUBRID

현재 벡터 검색 과제 연구/구현 현황

- 과제 수준의 연구 현황
  - 디스크 기반 HNSW 그래프 인덱스 구현
  - 벡터 검색 관련 최신 논문 서베이
    - B+ Tree 처럼 범용적으로 사용할만한 winner 인덱스 아직 없음
    - Row-based page 에서 locality 고려
    - 아직은 벡터 데이터/인덱스 도입과 관련한 일반적인 부분의 성능 고려 집중 필요
  - 연구 수준의 해볼 수 있는 최대한 모든 것 고려 중 (pgvector 이기기)
    - 기존의 다른 기능/쿼리와의 정합성 고려하지 않음

벡터 검색 기능은 CUBRID 발전 방향과 정렬되어 도입해야 한다

개발 목표/고려 사항: 벡터 검색 기능의 도입은 과제 종료 또는 그 이후

- 벡터 데이터: 사용자의 조회 대상이 아니지만, 크기가 크다
  - [PK | col1 | col2 | …. | vector (768 floats, 3K)]
    - OOS 프로젝트 (Out-of-line Overflow Storage)
- 벡터 인덱스: 그래프 구조
  - 새로운 인덱스 타입 추가
  - 인덱스 비동기 (Online) 빌드
  - 비동기 I/O (io_uring) SSD 병렬 읽기 최적화
  - SIMD 모듈 - 거리 계산에 사용
- 데이터 마이그레이션: 규모가 커지면 불가피
  - CDC + Debezium (Apache 2.0) + (커스텀화된 오픈소스 벡터 DB)
  - 벡터 연산 전용 서버 모드
- GPU 사용 환경 여부
  - 내부에 벡터 임베딩 모델 장착
  - GPU 가속

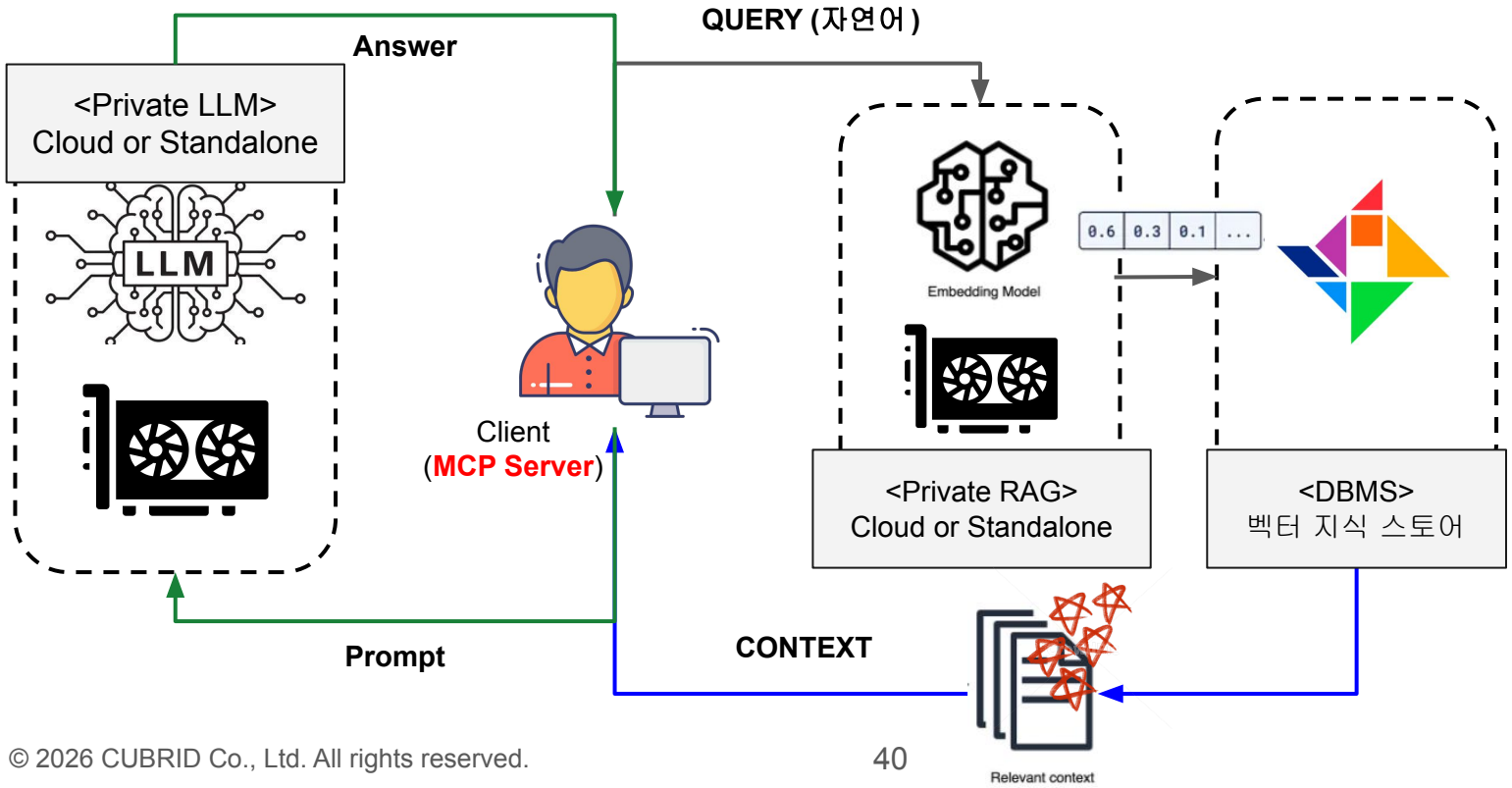
기업의 데이터를 반영한 결과의 **퀄리티**가 중요
- 정보의 정확성 (Accuracy)
- 정보의 최신성 (Freshness)

기업의 데이터에 대한 **보안**도 중요
- 프라이빗 데이터
- 데이터 접근 제한과 권한 관리
- 데이터 암호화와 로깅

스펙/기능적으로 어떻게 녹여 넣을 수 있을까?

- 임베딩 모델 장착 사례 참고:
  - https://www.oracle.com/kr/database/ai-vector-search/features/
  - https://www.tigerdata.com/blog/timescale-vector-x-llamaindex-making-postgresql-a-better-vector-database-for-ai-applications

다시, Private LLM 시스템 아키텍처에서



CUBRID

감사합니다