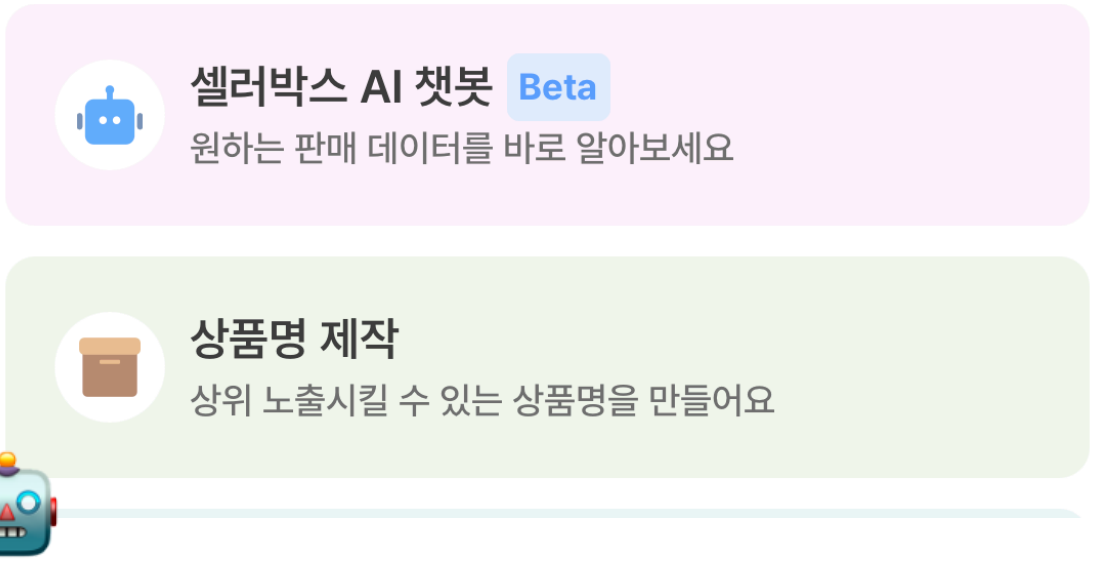


SQL Agent

태그	AI 데이터엔지니어 서클플랫폼 자연어처리
날짜	@2025년 2월 1일 → 2025년 3월 25일

개요

셀러박스에는 셀러분들의 많은 판매, 상품, 리뷰, 문의, 정산 데이터등을 가지고 있으며 기본적인 통계 대시보드외에도 사용자가 **궁금해할만한 질문을 DB 에서 직접 조회해서 찾고, 이를 통해 분석서비스를 제공하면 좋을 것 같다는 시도에서 시작**하였습니다.

역할

- SQL Agent 전체 챗봇구성

- SQL Agent 전용 DB 구성
- 챗봇 서빙 API 서버 및 인프라 구성

- 도전과제

☐ DB 내의 존재하는 민감 정보

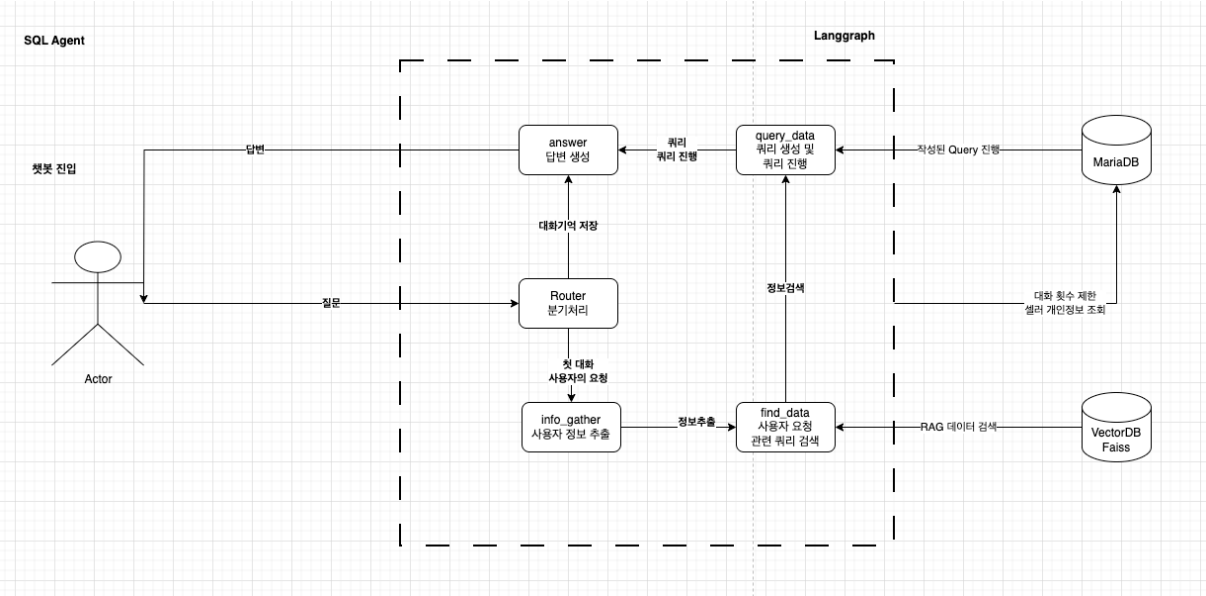
☐ 사용자의 자연어를 어떻게 적절한 SQL Query 로 변경할 것인가 (WHERE, JOIN 등)

구성

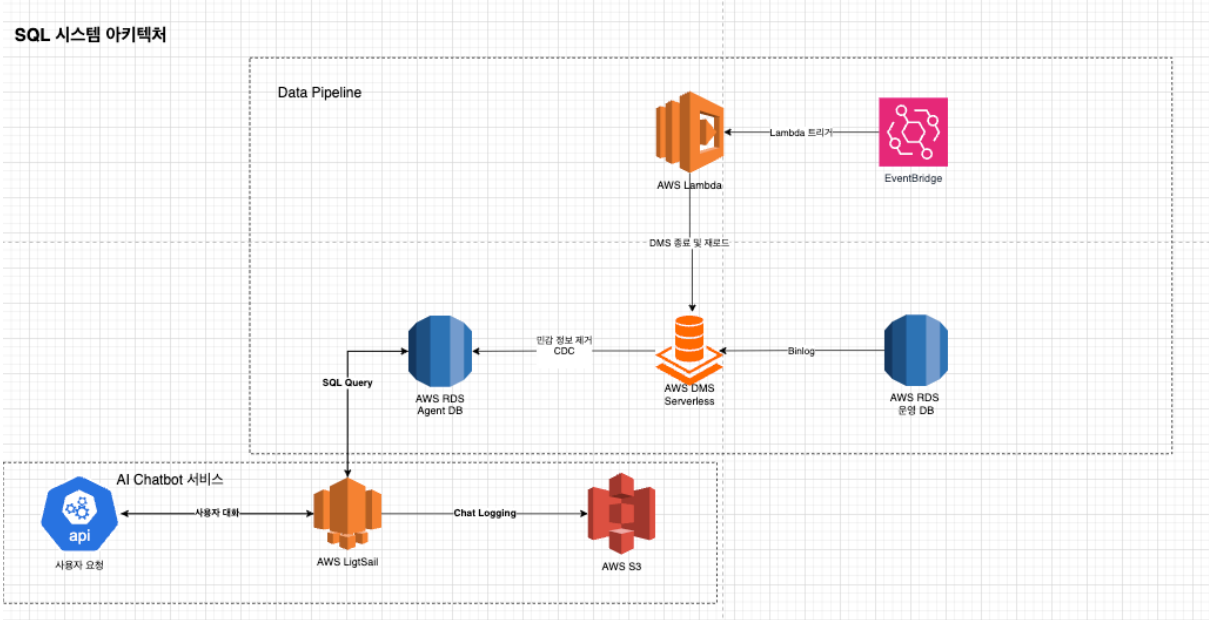
사용기술

- 서비스
  - AWS Lightsail
  - AWS RDS
  - AWS DMS
  - AWS Lambda
  - AWS EventBridge
- 언어
  - Python
  - SQL
- 프레임워크
  - FastAPI
- 라이브러리
  - Langchain
  - LangGraph
- DB
  - mysql
  - Faiss VectorDB

- LangGraph



- 시스템구성



실제 구현

☐ DB 내의 존재하는 민감 정보

👉 AWS DMS(Data Migration Service) 를 사용하여 SQL Agent 전용 DB 를 구성하여

☐ 사용자의 자연어를 어떻게 적절한 SQL Query 로 변경할 것인가 (WHERE, JOIN 등)

- RAG

실제 내부데이터로 기획자분들이나 디자이너분들이 어떤 데이터가 필요하다고 요청을 주시는데요.

👉 요청: 자연어, 응답: Query 및 데이터 로 전달드리는 경우가 많았습니다. **이에 착안하여 역으로 Query 를 만든 후 해당 쿼리에 대응되는 자연어를 작성 및 임베딩하여 VectorDB(Faiss)에 저장하여 SQL Agent 가 Query 를 구성하는데 참고하도록 하여 성능**을 높였습니다.

ex) 리뷰 관련 예시 쿼리

```
{
  "input_text":"가장 최근에 등록된 리뷰 10개를 가져와주세요",
  "query":"SELECT pr.product_name, pr.content, pr.score, pr.review
_date FROM review WHERE ..."
}
```

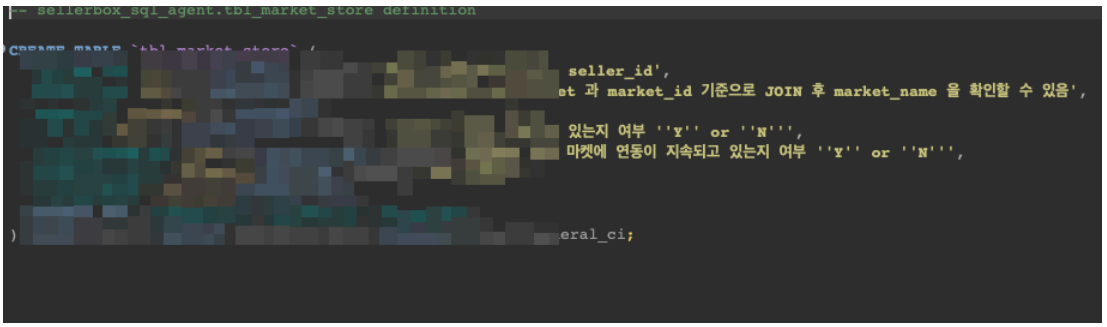
- DDL

또한 SQL Agent 는 내부적으로 자연어를 쿼리로 변경하면서 전체 테이블에서 어떤 테이블을 사용할지 RAG Query 리스트를 통해 판단하고 선정된 테이블들의 DDL 을 가져옵니다.

👉 이 때 가장 중요한 것은 DDL 의 코멘트 입니다.

- 각 컬럼이 어떤 의미를 가지고 있는지
- JOIN 할 수 있는 테이블은 어떤 것이 있는지 등등..

ex)

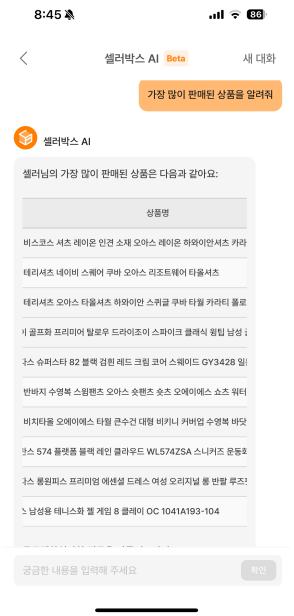


- Prompt

> 가장 중요한 요소 중 하나 였습니다.
>
> Agent 내부의 Node 마다 Instruction 과 Gudeline 등을 어떻게 구성하냐에 따라 퀄리티가 많이 바뀌었습니다.
>
> 👉 그러나 프롬프트의 영역은 100% 라는 것이 없는 것이 제가 개발하면서 얻은 결론이기 때문에
>
> example 을 만들고, OutPutParser 등을 사용해 강제할 수 있는 부분은 강제하는 방식으로 진행 하였습니다.

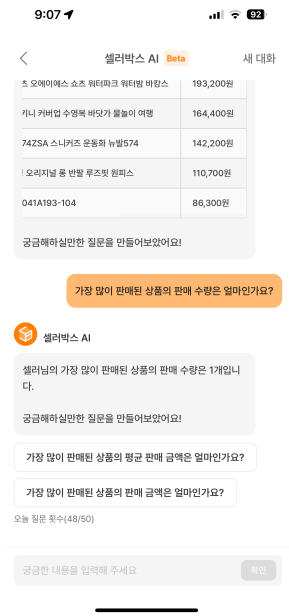
구현 화면

- 질문 1) 가장 많이 판매한 상품 알려줘
- 화면



- 내부 로직 input: 가장 많이 판매한 상품 알려줘
  - VectorDB 에서 검색
  - 증강된 데이터로 사용할 테이블 선정
  - 테이블 DDL 및 RAG 의 쿼리를 종합하여 Query 생성
  - 분석 제공

- 질문 2) Query 된 데이터 내부 질문 : 가장 많이 판매된 상품의 판매 수량은 얼마인가요?



- 내부 로직 input: 가장 많이 판매된 상품의 판매 수량은 얼마인가요?
  - 내부 Query 데이터를 통해 해당 질문에 답변할 수 있을지 판단
  - 가능
    - 내부 데이터에서 찾아서 답변 제공
  - 불가능
    - 사용자에게 불가능을 안내하여 이후 안내 [초기화, 다른질문]

회고

LangGraph 에서 제공하는 SQL Agent 문서를 기반으로 처음에는 쉽겠다 생각하여 진행하였으나

Tutorial 에서 제공되는 것은 SQL Agent 에 최적화된 DB 구조, DDL 가 이미 설계되어 있어 가능한 것이었습니다.

실제 운영 DB 는 운영중에 테이블이 추가되고 컬럼이 추가되는 등의 작업으로 해당 DB 를 직접 오래 다뤄본사람이 아니면 **쉽게 JOIN 을 하거나 원하는 데이터를 추출하는 과정이 어렵다는 것**을 알았고

이 어려운 과정은 SQL Agent 에게도 마찬가지였습니다.

👉 그래서 **해당 DB 를 처음보는 SQL 숙련자도 빠르게 요구사항을 뽑을 수 있을 만한 가이드를 Agent 에게 제공한다고 생각**하였고, 여러 자료등을 참고하여 **RAG 와 DDL 등의 코멘트를 적절하게 수정하는 방향으로 출시가능한 정도의 서비스**를 만들 수 있었습니다.

특히 AI 챗봇 서비스를 구성하면서 느낀 것은 ETL 파이프라인과 유사하다는 것을 알게되었습니다. 데이터를 모아서 재조립하고 하나의 데이터셋을 만드는 과정을 거치면서 결국은 모두 데이터가 이동하면서 의사 결정이 이뤄지는 과정이었고, 빠르게 개발할 수 있었던것도 파이프라인에 대한 이해가 있어서 그렇지 않을까 생각합니다.