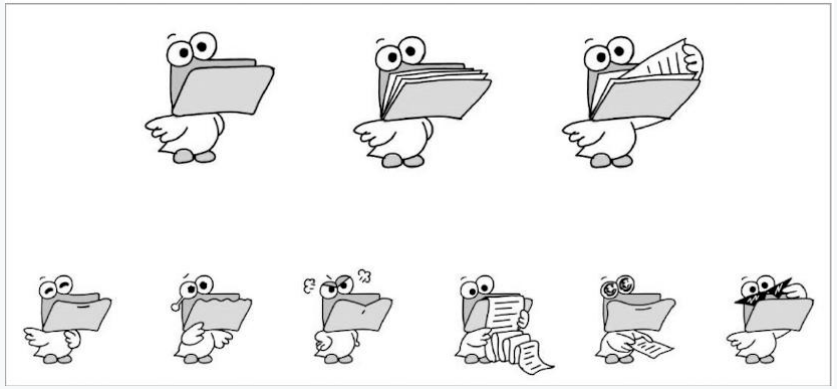
AI 데이터 분석가 '물어보새' 등장 2부

데이터 디스커버리

#AI #Artificial Intelligence #GenAI #RAG



2024. 08. 02

발표자: 태영

도입 및 배경

데이터 리터러시 향상과 AI 데이터 분석가 '물어보새'의 개발 계기, 목적, 핵심 기술(LLMOps, RAG, Text-to-SQL) 소개.

- 구성원의 데이터 활용 역량 향상을 위한 AI 도구 필요성
- 1편에서 다룬 물어보새의 핵심 기능과 기술
- 데이터 디스커버리 영역으로의 확장

기존 1편 요약

1편에서는 구성원의 데이터 리터러시 향상을 돕는 AI 데이터 분석가 '물어보새'의 기반 기술과 핵심 기능을 다루었습니다.

- 개발 계기와 목적: 구성원의 데이터 리터러시 향상 및 데이터 기반 의사결정 지원
- 핵심 기능: Text-to-SQL 기반 쿼리 생성 및 데이터 분석 지원
- 기술적 구현: RAG(검색 증강 생성), LLMOps(대규모 언어 모델 운영), Text-to-SQL 기술 활용
- 데이터 활용 방식: 비즈니스 문제 해결을 위한 쿼리 생성 중심 기능

새로운 확장: 데이터 디스커버리

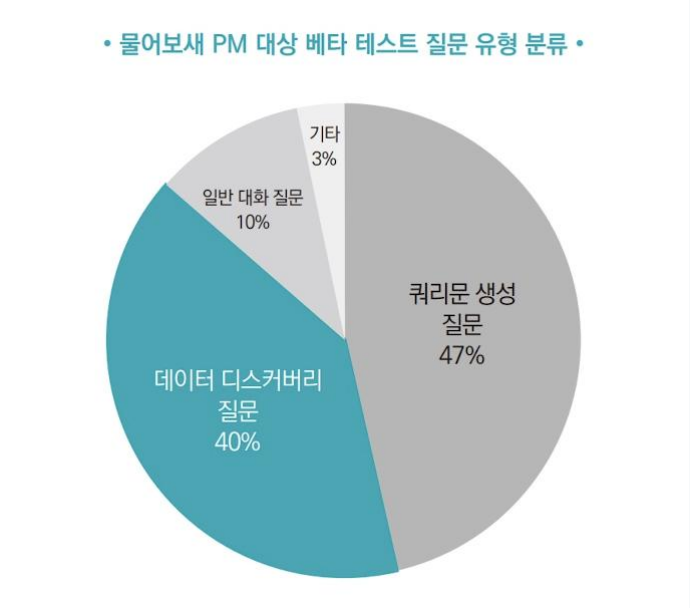
Text-to-SQL을 넘어 LLM으로 다양한 사내 데이터를 탐색, 인사이트 도출까지 지원하는 데이터 디스커버리 영역 확장

- 데이터 탐색 범위 확장: 사내 모든 테이블과 칼람, 로그 데이터 포함
- 데이터 이해 수준 향상: 쿼리문 해설, 테이블 설명, 활용 안내까지
- 데이터 활용 확장: 다양한 직군과 기술 수준의 구성원 모두 지원
- 데이터 플랫폼 연계: 데이터 카탈로그, 로그 체커와 연동 시너지

베타 테스트에서 얻은 인사이트

첫 번째 베타 테스트 (데이터 분석가/엔지니어 대상)

- 테이블 선별 정확도와 비즈니스 로직 반영 등 쿼리 생성 피드백 수집
- 트리노(Trino) 쿼리 함수 및 응답 시간 개선 필요성 확인
- 쿼리 생성뿐 아니라 쿼리 해설에 대한 요구 확인



두 번째 베타 테스트 (PM 대상)

- 물어보새의 실무적 활용성 및 비개발 직군 사용성 검증
- 다양한 형태의 데이터 디스커버리 질문 수집 (쿼리문 해설, 테이블 해설, 칼람 탐색, 로그 데이터 활용 등)
- 데이터 분석 직군 외 사용자를 위한 추가 기능 요구사항 파악

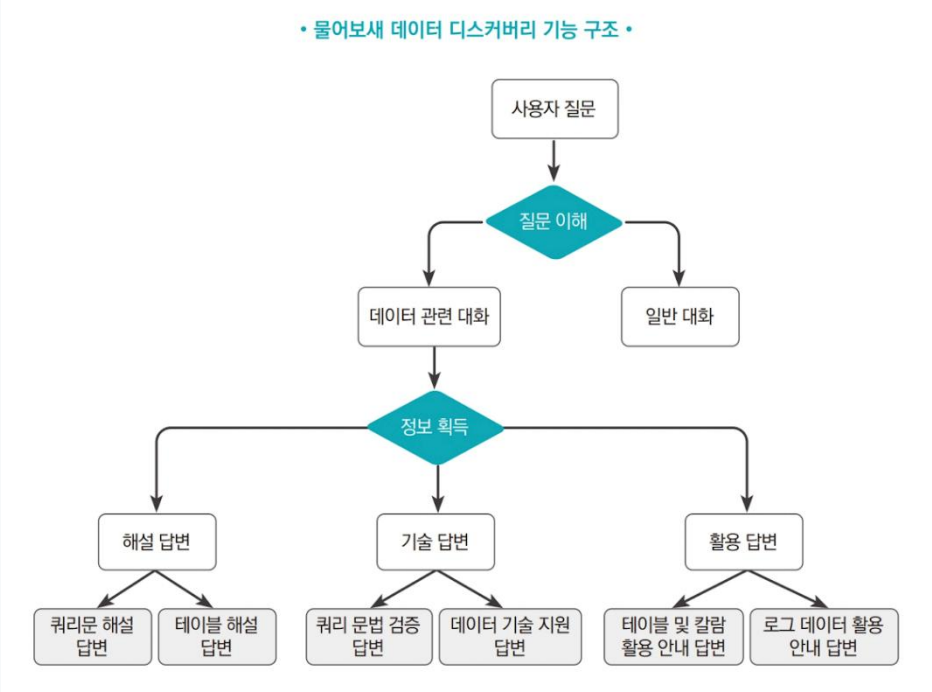
데이터 디스커버리 기능 개요

사용자 베타 테스트에서 얻은 인사이트를 바탕으로 사내 데이터 플랫폼과 연동하여 다양한질문을 이해하고 답변하는 데이터 디스커버리 기능 추가

- 데이터 카탈로그**(Data Catalog)**: 사내 모든 테이블과 칼럼에 대한 상세 정보 저장
- 로그 체커**(Log Checker)**: 앱과 웹에서 발생하는 사용자 행동과 이벤트 정보를 로그 단위로 통합 관리
- 벡터 스토어 활용: 사용자 질문의 의도를 파악하고 관련 정보를 찾아 답변 제공

> → 쿼리문/테이블 해설, 칼럼 안내, 로그 활용 등 다양한 정보 획득 기능으로 데이터 활용 시너지 극대화

데이터 디스커버리 계층 구조



- 질문 이해 – 정보 획득 순으로 계층화된 구조 설계
- 라우터 슈퍼바이저 체인 이용 - 사용자 질문의 의도를 정확히 파악하고 적절한 답변 전달
- 계층 구조를 통해 **LLM** 환각 문제 최소화와 확장성 확보 가능

질문 이해 단계 설계

목표: 사용자 질문의 의도와 수준을 정확히 파악하여 답변 품질을 높이고 완성도 높은 질문으로 연결

- 라우터 슈퍼바이저 체인: 질문 의도를 파악하고 적절한 답변을 전달하기 위한 구조 구현
- 질문 분류 시스템: 질문을 데이터/비즈니스 관련 여부로 분류 후, 정보 획득 유형별 추가 분류
- 질문 품질 평가: 프롬프트 엔지니어링 기법으로 질문의 완성도와 품질을 점수화
- 질문 개선 피드백: 품질이 낮은 질문에 대해 개선 방향과 구체적 예시 제공

질문 평가와 개선

프롬프트 엔지니어링을 활용한 질문 품질 향상과 사용자 역량 강화

질문 해석 능력 개선

- 질문 평가 기준 수립 및 점수화 시스템
- 프롬프트 엔지니어링 기법으로 일관된 평가
- 벡터 스토어로 사내 용어·질문 결합
- 추상적/전문적 질문을 이해하기 쉽게 변환

질문 생성 능력 개선

- 질문 기준 미달 시 자동 피드백 제공
- "더 구체적으로 질문하세요" 등 안내 생성
- 적합한 질문 예시로 가이드 제공
- 슬랙앱 튜토리얼 및 활용 안내 화면 개발

대화 유형별 처리 방식

물어보새는 다양한 대화 유형에 따라 처리 방식을 최적화하여 적용합니다.

싱글턴(Single-Turn)

하나의 질문과 응답으로 구성. 구축이 간단하고 응답 속도가 빠르지만 문맥이 유지되지 않음. 질문 이해 단계에서 주로 사용

가이디드 싱글턴

하나의 질문과 응답이지만 특정 방향으로 대화 유도. 질문이 구체적이지 않을 때 질문 작성 가이드 제공

멀티턴(Multi-Turn)

질문과 응답이 연속적으로 이어지며 문맥 유지. 긴 대화와 지속적인 상호작용이 가능하지만 허위 생성 가능성 있음

- 자동 분류 시스템 : LLM 기반 프롬프트 엔지니어링으로 질문 유형에 따라 분기 처리
- 비즈니스 관련 여부 : 데이터 또는 비즈니스 관련 질문으로 분류 후 정보 획득 유형 선택
- 현재 개발 중 : 안정적인 서비스 제공이 가능한 멀티턴 기능 지속 개발

정보 획득 단계 설계

질문 이해 단계에서 분류된 사용자 질문에 구체적인 답변을 제공하는 단계

1 쿼리문/테이블 해설 - 복잡한 쿼리문과 테이블 구조를 자연어로 해석

2 쿼리 문법 검증, 기술 지원 - 칼럼명 오류 보정, 실행 최적화 제안

3 테이블·칼럼 활용 안내 - 특정 정보가 담긴 테이블/칼럼 탐색 지원

4 로그 데이터 활용 안내 - 로그 체커 기반 관련 로그 정보 제공

* 각 기능별로 맞춤형 체인과 알고리즘을 구현하여 정확한 답변 제공

쿼리문 및 테이블 해설 기능

쿼리문 내 주요 조건, 칼람, 결과값 등 자연어 해설 제공

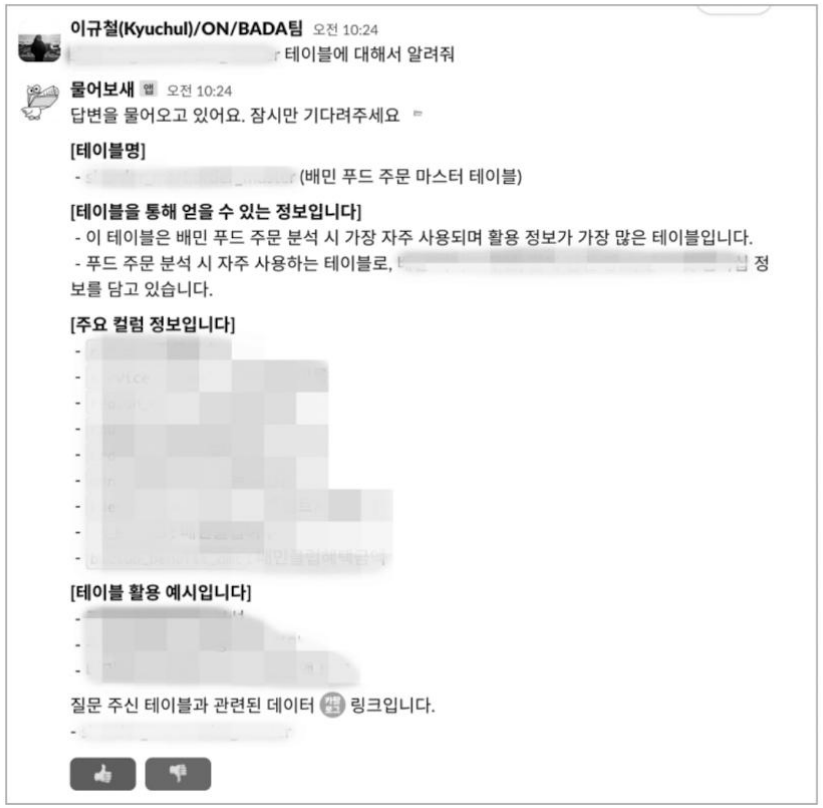
- 구현 방법: SQLGlot과 DDL 벡터스토어를 활용하여 쿼리문과 테이블의 정보를 추출
- 고도화 기법: Plan and Solve Prompting 적용으로 쿼리문과 테이블에 대한 해석 품질 향상
- 추가 기능: 데이터 카탈로그 정보 링크 제공으로 상세 정보 탐색 지원

SQLGlot

SQLGlot은 SQL 파서 및 변환기로, 다양한 SQL 방언 간 변환, 문법 검증, 최적화, 쿼리 해석 등에 사용

쿼리를 AST(Abstract Syntax Tree, 추상 구문 트리) 형태로 변환하여 SQL 문장의 구조를 계층적으로 분석하고 처리

• 물어보새 테이블 해설 답변 예시 •



쿼리 문법 검증 및 데이터 기술 지원

쿼리문 문법 검증 기능은 두 단계의 세부 체인으로 구성되어 허위 생성 가능성 최소화 및 성능 향상을 실현합니다.

칼럼명 보정 체인

- 쿼리문에서 테이블명과 칼럼명 추출
- DDL 기반 칼럼명 오류 확인 및 보정
- DDL 축소 후 다음 단계로 전달

문법 검증 및 최적화 체인

- 보정된 쿼리와 축소된 DDL 활용
- 문법, 칼람 값 오류 확인
- 쿼리 실행 최적화 방안 제안

- 데이터 기술 지원: 쿼리 함수 및 데이터베이스 관련 전문 지식 제공
- 향후 계획: 비즈니스 로직 메타 정보 구축 및 사용자 가이드 개선

테이블 및 칼람 활용 안내

사용자가 필요한 데이터를 더 쉽게 찾고 활용할 수 있도록 특정 정보를 담고 있는 테이블명과 칼람 정보를 제공하는 기능입니다.

- 메타데이터 고도화 - LLM을 이용해 테이블의 목적, 특성, 주요 키워드 등 메타데이터 구축
- 질문 구체화 체인 - 비즈니스 용어 사전과 토픽 모델링을 활용해 사용자 질문 확장 및 키워드 선별
- 혼합 검색 체인 - 테이블 메타데이터와 DDL을 이용한 리트리버와 LLM의 3단계 검색으로 최적 테이블 선별
- 사용자 인터페이스 - 테이블명, 주요 칼람, 활용 예시 정보와 함께 데이터 카탈로그 링크 제공

※ 향후 개선 계획: 메타데이터 생성 프롬프트 고도화와 보완 로직으로 허위 생성 문제 해결

로그 데이터 활용 안내

로그 체커 데이터 기반의 비정형 로그 데이터 탐색 및 활용 기능

• 로그 체커 데이터 예시 •

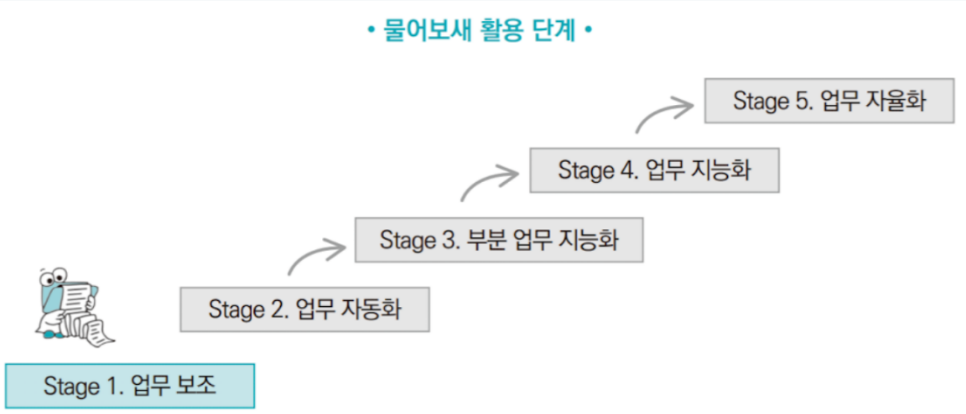
로그 이름	로그 설명	Screen Name	Group	Event	Type	새롭게 정의된 로그 설명
가게 상세 〉 쿠폰 받기	배포 일자 기록	ShopDet	Cpn	CpnDown	Click	가게 상세 쿠폰 다운로드 클릭
1배민스토어 〉 가게카드 노출	배포 앱 버전 기록	Store	SellerCard	SellerCard	Imp	배민스토어 셀러 카드 노출

- 로그 데이터 해석의 과제 : '로그 이름'의 정보 부족, 로그 체커의 고유 조합값(Screen Name, Group, Event, Type) 활용 필요
- 주요 구현 기법 : 로그 용어 사전 구축(영-한 번역), 로그 용어보정 사전 추가(회사 특화 용어), 축약어 유사도 기반 매핑 ('ShopDet'→'Shop Detail')
- 성과 : LLM으로 복잡한 검색 알고리즘 대체, 정확한 로그 정보 제공으로 다양한 사용자의 로그 이해도 향상

물어보새 발전 계획 및 로드맵

물어보새는 다음과 같은 다단계 업무지능화 로드맵을 통해 지속적으로 발전해 나갈 계획입니다.

- 지식 생성 단계: 데이터 기반 인사이트 자동 생성 및 지식화 시스템 구축
- **AI** 에이전트 확장: 비즈니스 도메인 전문 에이전트로 진화, 협업 시스템 구현
- **BI** 포털/대시보드 연동: 다양한 데이터 플랫폼과 연계, 시각화 자동화 구현
- **BADA**팀 신설: 비즈니스 AI와 데이터 활용 시너지 확대를 위한 조직 개편



서비스 단계	1. 업무 보조	2. 업무 자동화	3. 부분 업무 지능화	4. 업무 지능화	5. 업무 자율화
의사 결정 담당	구성원	구성원	구성원	구성원 + 물어보새	물어보새
데이터 리터러시	데이터 이해 및 추출 지원	데이터 생성 및 공유	데이터 분석 업무 수행	데이터 기반 의사 결정 지원	데이터 기반 의사 결정 최적화

Q&A 및 종료

질문을 환영합니다