**1. Nhập thông tin và thiết lập kết nối:**

* **Giao diện người dùng (React):**
  + Người dùng nhập các thông tin kết nối như: địa chỉ IP máy chủ, tên cơ sở dữ liệu, tên người dùng, mật khẩu và chọn mô hình LLM.
  + Khi người dùng nhấn "Kết nối," các thông tin này sẽ được gửi tới backend qua một API call.

**2. Thu thập metadata qua Spring Boot:**

* **Backend Spring Boot:**
  + Nhận thông tin kết nối từ giao diện React.
  + Kiểm tra tính hợp lệ của dữ liệu đầu vào để đảm bảo tất cả các trường bắt buộc đều có.
  + Thiết lập kết nối với cơ sở dữ liệu được chỉ định bằng thông tin đăng nhập đã cung cấp.
  + Sau khi kết nối thành công, backend sẽ chạy script DataHub để thu thập metadata từ cơ sở dữ liệu (các bảng, cột, mối quan hệ, v.v.).
  + Khi quá trình thu thập metadata thành công, backend sẽ gửi thông báo thành công về cho giao diện React.

**3. Thông báo cho người dùng:**

* **Giao diện người dùng (React):**
  + Thông báo thành công từ backend sẽ kích hoạt một cập nhật trên giao diện, thông báo cho người dùng rằng quá trình thu thập metadata đã thành công và họ có thể tiếp tục thực hiện truy vấn cơ sở dữ liệu.

**4. Quá trình thực hiện truy vấn:**

* **Tương tác của người dùng:**
  + Người dùng tương tác với giao diện chatbot để đặt câu hỏi bằng ngôn ngữ tự nhiên.
  + Câu hỏi của người dùng sẽ được gửi đến backend qua một API call khác.

**5. Tạo và thực thi truy vấn:**

* **Backend Spring Boot:**
  + Nhận truy vấn của người dùng.
  + Chuyển truy vấn này tới mô hình LLM đã chọn để tạo ra truy vấn SQL tương ứng.
  + Truy vấn SQL được tạo ra sẽ được thực thi trên cơ sở dữ liệu.
  + Kết quả của truy vấn SQL sau đó sẽ được xử lý và chuẩn bị để hiển thị.

**6. Trả kết quả về cho người dùng:**

* **Giao diện người dùng (React):**
  + Kết quả truy vấn đã được xử lý sẽ được gửi lại cho giao diện React.
  + Kết quả sẽ được hiển thị dưới dạng dễ hiểu trên giao diện, hoàn thành chu trình truy vấn của người dùng.

**7. Xử lý lỗi:**

* Nếu bất kỳ bước nào gặp lỗi (ví dụ: kết nối thất bại, thu thập metadata gặp vấn đề, lỗi khi tạo truy vấn), một thông báo lỗi sẽ được gửi lại cho giao diện React để thông báo cho người dùng.

**8. Tương tác liên tục:**

* Người dùng có thể tiếp tục đặt câu hỏi, và chu trình tạo truy vấn, thực thi và trả kết quả sẽ lặp lại khi cần thiết.

**1. Thành phần Phân tích Đầu vào:**

* **Mục tiêu:** Phân tích câu truy vấn ngôn ngữ tự nhiên và trích xuất các đặc điểm giúp xác định kỹ thuật RAG phù hợp.
* **Các đặc điểm cần trích xuất:**
  + **Độ phức tạp của truy vấn:** Xác định xem truy vấn là đơn giản (ví dụ: câu lệnh SELECT cơ bản) hay phức tạp (ví dụ: liên quan đến nhiều JOIN, subqueries, v.v.).
  + **Loại truy vấn:** Phân loại truy vấn (ví dụ: truy vấn lấy dữ liệu, tổng hợp, lọc).
  + **Nhận biết về cấu trúc dữ liệu:** Kiểm tra xem truy vấn có đề cập rõ ràng đến tên bảng, cột, hoặc các thực thể liên quan đến cấu trúc dữ liệu hay không.
  + **Thuật ngữ đặc thù trong lĩnh vực:** Xác định xem truy vấn có sử dụng ngôn ngữ đặc thù của lĩnh vực cụ thể không, có thể yêu cầu tìm kiếm bổ sung ngữ cảnh.
* **Kỹ thuật sử dụng:**
  + Sử dụng kết hợp các kỹ thuật NLP như Nhận diện thực thể (NER) để nhận biết các thành phần cấu trúc dữ liệu và các thuật ngữ đặc thù.
  + Tận dụng các mô hình ngôn ngữ đã được huấn luyện để đánh giá độ phức tạp ngữ nghĩa của truy vấn.

**2. Bộ chọn Chiến lược RAG:**

* **Mục tiêu:** Dựa trên phân tích, quyết định nên sử dụng kỹ thuật RAG nào để tạo ra truy vấn SQL.
* **Tiêu chí quyết định:**
  + **Truy vấn đơn giản:** Nếu truy vấn đơn giản, sử dụng các kỹ thuật tìm kiếm cơ bản, như tìm kiếm thưa với dynamic prompting.
  + **Truy vấn phức tạp:** Đối với các truy vấn phức tạp hơn, chọn tìm kiếm lai (thưa + dày), các cơ chế cross-attention, hoặc tinh chỉnh lặp lại.
  + **Truy vấn đặc thù lĩnh vực:** Nếu truy vấn chứa thuật ngữ đặc thù trong lĩnh vực, tăng cường tìm kiếm với đồ thị tri thức hoặc embedding đặc thù trong lĩnh vực.
  + **Truy vấn phụ thuộc vào cấu trúc dữ liệu:** Nếu truy vấn phụ thuộc nhiều vào cấu trúc dữ liệu bên dưới, sử dụng các kỹ thuật kết hợp tìm kiếm cấu trúc dữ liệu, như tăng cường ngữ cảnh với đồ thị tri thức.
* **Triển khai:**
  + Có thể sử dụng hệ thống dựa trên cây quyết định hoặc luật để triển khai ban đầu.
  + Đối với các hệ thống tiên tiến hơn, có thể sử dụng mô hình học máy được huấn luyện trên dữ liệu đã được chú thích, trong đó các kỹ thuật RAG khác nhau đã được áp dụng.

**9. Các tình huống yêu cầu từ người dùng:**

**Truy vấn thông tin cơ bản:**

* **Lấy thông tin bảng:**
  + "Hiển thị tất cả các bảng trong cơ sở dữ liệu."
  + "Các cột trong bảng 'nhân viên' là gì?"
* **Truy xuất dữ liệu:**
  + "Lấy tất cả các bản ghi từ bảng 'bán hàng'."
  + "Liệt kê tên của tất cả các nhân viên gia nhập sau năm 2020."

**Lọc và sắp xếp dữ liệu:**

* **Lọc dữ liệu:**
  + "Hiển thị tất cả các đơn hàng của khách hàng 'John Doe'."
  + "Lấy tất cả các nhân viên có mức lương lớn hơn 50,000 USD."
* **Sắp xếp dữ liệu:**
  + "Liệt kê các sản phẩm theo giá, từ cao đến thấp."
  + "Hiển thị 5 giao dịch gần đây nhất."

**Tính toán tổng hợp và nhóm dữ liệu:**

* **Hàm tổng hợp:**
  + "Mức lương trung bình của nhân viên trong phòng IT là bao nhiêu?"
  + "Tổng doanh thu của tháng 6 năm 2024 là bao nhiêu?"
* **Nhóm dữ liệu:**
  + "Nhóm các nhân viên theo phòng ban và hiển thị số lượng trong mỗi phòng."
  + "Hiển thị tổng doanh thu được tạo ra bởi mỗi khu vực bán hàng."

**Kết nối các bảng:**

* **Kết nối đơn giản:**
  + "Lấy danh sách đơn hàng cùng với tên khách hàng."
  + "Hiển thị chi tiết sản phẩm cùng với thông tin nhà cung cấp."
* **Kết nối phức tạp:**
  + "Liệt kê tất cả các nhân viên và tên của quản lý của họ."
  + "Tìm các đơn hàng mà sản phẩm đã được vận chuyển sau ngày hạn cùng với tên khách hàng."

**Yêu cầu sửa đổi dữ liệu:**

* **Chèn dữ liệu:**
  + "Thêm một nhân viên mới vào bảng 'nhân viên'."
  + "Chèn một đơn hàng mới với các chi tiết sau..."
* **Cập nhật dữ liệu:**
  + "Cập nhật lương của 'John Doe' thành 60,000 USD."
  + "Thay đổi trạng thái của đơn hàng #12345 thành 'đã vận chuyển'."
* **Xóa dữ liệu:**
  + "Xóa tất cả các bản ghi từ bảng 'dữ liệu tạm thời'."
  + "Xóa nhân viên có ID 456 khỏi bảng 'nhân viên'."

**Truy vấn cấu trúc cơ sở dữ liệu:**

* **Thông tin lược đồ:**
  + "Lược đồ của bảng 'đơn hàng' là gì?"
  + "Hiển thị cấu trúc của cơ sở dữ liệu 'khách hàng'."
* **Ràng buộc và chỉ mục:**
  + "Liệt kê tất cả các khóa chính trong bảng 'sản phẩm'."
  + "Những chỉ mục nào có trong bảng 'giao dịch'?"

**Yêu cầu về hiệu suất và tối ưu hóa:**

* **Hiệu suất truy vấn:**
  + "Giải thích kế hoạch truy vấn để truy xuất tất cả các đơn hàng."
  + "Làm thế nào để tối ưu hóa truy vấn lấy top 10 khách hàng theo doanh thu?"
* **Khuyến nghị chỉ mục:**
  + "Gợi ý các chỉ mục để cải thiện hiệu suất của các truy vấn trên bảng 'bán hàng'."

**Metadata và thông tin hệ thống:**

* **Metadata cơ sở dữ liệu:**
  + "Hiển thị metadata đã thu thập cho bảng 'nhân viên'."
  + "Liệt kê tất cả các mối quan hệ giữa các bảng trong cơ sở dữ liệu này."
* **Thông tin hệ thống:**
  + "Phiên bản cơ sở dữ liệu hiện tại là gì?"
  + "Hiển thị số lượng kết nối đang hoạt động với cơ sở dữ liệu."

**Xuất nhập dữ liệu:**

* **Xuất dữ liệu:**
  + "Xuất bảng 'nhân viên' thành tệp CSV."
  + "Xuất tất cả dữ liệu từ bảng 'bán hàng' cho năm 2024."
* **Nhập dữ liệu:**
  + "Nhập dữ liệu từ 'data.csv' vào bảng 'khách hàng'."

**Xử lý lỗi và gỡ lỗi:**

* **Xử lý lỗi:**
  + "Có gì sai trong truy vấn cuối cùng của tôi?"
  + "Hiển thị nhật ký của lần thử kết nối cơ sở dữ liệu gần nhất."
* **Gỡ lỗi truy vấn:**
  + "Tại sao truy vấn của tôi không trả về kết quả nào?"
  + "Kiểm tra xem có lỗi cú pháp nào trong câu lệnh SQL của tôi không."

**Truy vấn phân tích nâng cao:**

* **Tính toán phức tạp:**
  + "Tính tăng trưởng doanh thu theo năm."
  + "Tìm mối tương quan giữa giá sản phẩm và khối lượng bán hàng."
* **Phân tích chuỗi thời gian:**
  + "Hiển thị xu hướng bán hàng trong 12 tháng qua."
  + "Tìm trung bình động của doanh thu hàng ngày trong 30 ngày qua."

**Thực thi SQL tùy chỉnh:**

* **Chạy SQL tùy chỉnh:**
  + "Thực thi truy vấn SQL tùy chỉnh sau..."
  + "Chạy thủ tục lưu trữ có tên 'UpdateInventory'."

**Quản lý người dùng và quyền truy cập:**

* **Quản lý người dùng:**
  + "Thêm một người dùng mới với quyền chỉ đọc."
  + "Thu hồi quyền truy cập của người dùng 'johndoe'."
* **Kiểm soát truy cập:**
  + "Cấp quyền select trên bảng 'bán hàng' cho vai trò 'analyst'."
  + "Liệt kê tất cả các người dùng có quyền ghi trên bảng 'nhân viên'."

**Ví dụ 1: Truy vấn Đơn Giản**

* **Truy vấn:** "Hiển thị danh sách tất cả nhân viên trong công ty."
* **Phân Tích:**
  + **Độ phức tạp:** Đơn giản, chỉ cần một câu lệnh SELECT cơ bản mà không có điều kiện hoặc kết nối phức tạp.
  + **Ngữ cảnh:** Không yêu cầu thông tin về cấu trúc dữ liệu.
* **Kỹ Thuật RAG Được Chọn:** **Dynamic Prompting with Retrieval (Nhắc nhở động với truy xuất)**
* **Lý Do:** Sử dụng kỹ thuật này vì nó có thể lấy các ví dụ tương tự từ cơ sở dữ liệu của các truy vấn trước đó như "Hiển thị danh sách tất cả khách hàng" hoặc "Danh sách tất cả sản phẩm," và sử dụng những ví dụ này để tạo ra câu lệnh SQL phù hợp mà không cần phức tạp hóa.

**Ví dụ 2: Truy vấn Phức Tạp với Nhiều Join**

* **Truy vấn:** "Tìm tên các khách hàng đã đặt mua sản phẩm thuộc danh mục 'Điện tử' và chi tiêu hơn 500 đô la trong năm ngoái."
* **Phân Tích:**
  + **Độ phức tạp:** Cao, yêu cầu nhiều điều kiện và kết nối giữa các bảng (vd: bảng customers, orders, và products).
  + **Ngữ cảnh:** Cần phải hiểu mối quan hệ giữa các bảng.
* **Kỹ Thuật RAG Được Chọn:** **Sparse-Dense Hybrid Retrieval (Kết hợp truy xuất thưa và dày đặc)**
* **Lý Do:** Kỹ thuật này đảm bảo rằng hệ thống có thể lấy lại các ví dụ có liên quan dựa trên cả từ khóa và ngữ nghĩa, giúp tạo ra câu lệnh SQL chính xác hơn trong các tình huống phức tạp.

**Ví dụ 3: Truy vấn Phụ Thuộc Vào Sơ Đồ Dữ Liệu**

* **Truy vấn:** "Liệt kê tất cả các đơn hàng đặt trong tháng vừa qua cùng với email của khách hàng."
* **Phân Tích:**
  + **Độ phức tạp:** Trung bình, yêu cầu hiểu biết về các cột và mối quan hệ giữa các bảng.
  + **Ngữ cảnh:** Cần truy xuất thông tin về sơ đồ dữ liệu.
* **Kỹ Thuật RAG Được Chọn:** **Contextual Augmentation with Knowledge Graphs (Bổ sung ngữ cảnh với đồ thị tri thức)**
* **Lý Do:** Đồ thị tri thức có thể cung cấp thông tin về sơ đồ dữ liệu cần thiết, như mối quan hệ giữa các bảng, để hệ thống tạo ra câu lệnh SQL phản ánh đúng cấu trúc dữ liệu cơ bản.

**Ví dụ 4: Truy vấn Theo Ngữ Cảnh Đặc Thù**

* **Truy vấn:** "Lấy danh sách tất cả các bệnh nhân đã được chẩn đoán mắc bệnh tiểu đường và được kê đơn Metformin trong năm ngoái."
* **Phân Tích:**
  + **Độ phức tạp:** Cao, sử dụng thuật ngữ chuyên ngành và có thể yêu cầu kết nối phức tạp giữa các bảng liên quan đến y tế.
  + **Ngữ cảnh:** Yêu cầu sự hiểu biết sâu về thuật ngữ y tế.
* **Kỹ Thuật RAG Được Chọn:** **Cross-Attention Mechanisms (Cơ chế chú ý chéo)**
* **Lý Do:** Cơ chế chú ý chéo giúp mô hình hiểu rõ hơn về cả truy vấn ngôn ngữ tự nhiên và ngữ cảnh được truy xuất, đảm bảo rằng câu lệnh SQL được tạo ra chính xác với sự hiểu biết về thuật ngữ chuyên ngành.

**Ví dụ 5: Truy vấn Mơ Hồ hoặc Không Chắc Chắn**

* **Truy vấn:** "Cho tôi bản tóm tắt về doanh số bán hàng của quý gần nhất."
* **Phân Tích:**
  + **Độ phức tạp:** Mơ hồ, thuật ngữ "tóm tắt" và "quý gần nhất" có thể có nhiều cách hiểu khác nhau.
  + **Ngữ cảnh:** Cần làm rõ ngữ cảnh hoặc thuật ngữ.
* **Kỹ Thuật RAG Được Chọn:** **Iterative Refinement with RAG (RAG với tinh chỉnh lặp)**
* **Lý Do:** Kỹ thuật tinh chỉnh lặp cho phép hệ thống tạo ra câu lệnh SQL ban đầu và sau đó tinh chỉnh dựa trên ngữ cảnh hoặc phản hồi từ truy xuất, giúp đảm bảo câu lệnh cuối cùng phản ánh đúng ý định của người dùng.

**Ví dụ 6: Truy vấn Phức Tạp với Học Tập Ít Ví Dụ**

* **Truy vấn:** "Tính tổng doanh thu của từng danh mục sản phẩm trong năm tài chính vừa qua."
* **Phân Tích:**
  + **Độ phức tạp:** Cao, bao gồm các phép tính tổng hợp (SUM doanh thu) và nhóm (GROUP BY danh mục sản phẩm).
  + **Ngữ cảnh:** Yêu cầu ví dụ phức tạp và thích ứng.
* **Kỹ Thuật RAG Được Chọn:** **Few-Shot Learning with Adaptive Retrieval (Học ít ví dụ với truy xuất thích ứng)**
* **Lý Do:** Học ít ví dụ cho phép mô hình được cung cấp một số ví dụ liên quan (được truy xuất dựa trên độ phức tạp của truy vấn) làm một phần của prompt, giúp tạo ra câu lệnh SQL chính xác hơn.

**Ví dụ 7: Truy vấn Thường Xuyên với Tần Suất Cao**

* **Truy vấn:** "Tổng số người dùng đang hoạt động là bao nhiêu?"
* **Phân Tích:**
  + **Độ phức tạp:** Thấp, nhưng có thể được hỏi thường xuyên.
  + **Ngữ cảnh:** Có thể sử dụng câu lệnh SQL mẫu có sẵn.
* **Kỹ Thuật RAG Được Chọn:** **Pre-computed Templates with RAG (Mẫu đã tính toán trước với RAG)**
* **Lý Do:** Đối với các truy vấn thường xuyên, một mẫu câu lệnh SQL hoặc câu lệnh SQL đã được tính toán trước có thể được sử dụng, giảm thiểu thời gian xử lý và đảm bảo phản hồi nhanh chóng.

**Ví dụ 8: Truy vấn Ngôn Ngữ Tự Nhiên với Chú Thích Sơ Đồ Dữ Liệu**

* **Truy vấn:** "Tìm mức lương trung bình của nhân viên trong phòng Kỹ thuật."
* **Phân Tích:**
  + **Độ phức tạp:** Trung bình, yêu cầu hiểu về cấu trúc sơ đồ dữ liệu và các cột liên quan.
  + **Ngữ cảnh:** Cần chú thích và hiểu biết về sơ đồ dữ liệu.
* **Kỹ Thuật RAG Được Chọn:** **Dual-Encoder Models (Mô hình mã hóa kép)**
* **Lý Do:** Mô hình mã hóa kép có thể mã hóa cả truy vấn ngôn ngữ tự nhiên và sơ đồ dữ liệu, cho phép truy xuất các mẫu SQL liên quan hoặc tạo SQL trực tiếp dựa trên chú thích sơ đồ.

**Ví dụ 9: Truy vấn với Tổng Hợp và Lọc Phức Tạp**

* **Truy vấn:** "Hiển thị 10 sản phẩm có doanh thu cao nhất theo quý trong hai năm vừa qua."
* **Phân Tích:**
  + **Độ phức tạp:** Cao, bao gồm các phép tổng hợp, nhóm và lọc theo thời gian.
  + **Ngữ cảnh:** Yêu cầu kỹ thuật tinh chỉnh và truy xuất.
* **Kỹ Thuật RAG Được Chọn:** **Fine-Tuning with Retrieval-Augmented Data (Tinh chỉnh với dữ liệu tăng cường truy xuất)**
* **Lý Do:** Tinh chỉnh mô hình LLM trên tập dữ liệu bao gồm các ví dụ với ngữ cảnh truy xuất có thể giúp mô hình hiểu và tạo ra các câu lệnh SQL phức tạp một cách chính xác hơn.

**Ví dụ 10: Truy vấn với Cấu Trúc Biến Động Cao**

* **Truy vấn:** "Cung cấp bản phân tích đơn hàng theo quốc gia và phương thức thanh toán."
* **Phân Tích:**
  + **Độ phức tạp:** Trung bình đến cao, với nhiều cách hiểu khác nhau dựa trên cấu trúc dữ liệu.
  + **Ngữ cảnh:** Có thể yêu cầu nhiều kỹ thuật RAG khác nhau để xử lý.
* **Kỹ Thuật RAG Được Chọn:** **Model Ensemble with RAG (Tập hợp mô hình với RAG)**
* **Lý Do:** Sử dụng kỹ thuật tập hợp mô hình cho phép nhiều kỹ thuật RAG được áp dụng đồng thời, đảm bảo rằng câu lệnh SQL cuối cùng được chọn dựa trên độ tin cậy và sự phù hợp của các mô hình.