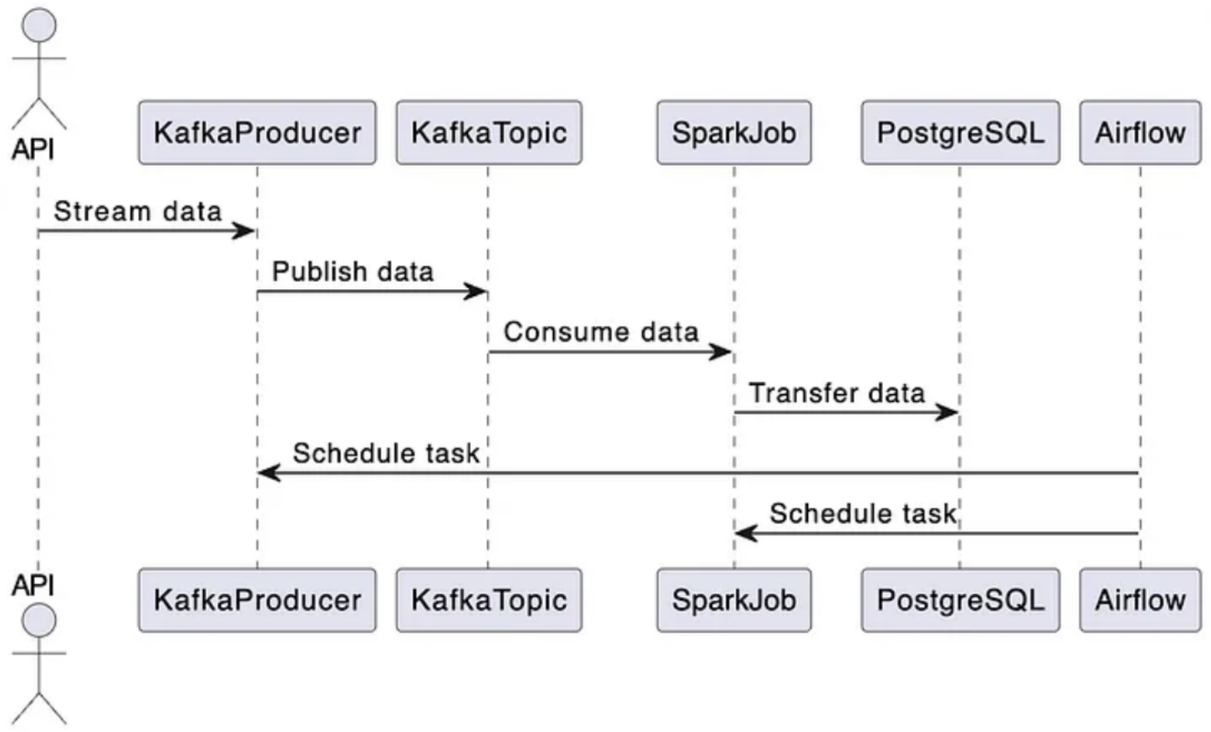

Truyền dữ liệu: Ban đầu, dữ liệu được truyền từ API vào Kafka topic

Xử lý dữ liệu: Sau đó, một Spark Job sẽ tiếp quản, tiêu thụ dữ liệu từ Kafka topic và chuyển nó sang cơ sở dữ liệu PostgreSQL

Lập lịch với Airflow: Cả tác vụ phát trực tuyến và Spark Job đề được sắp xếp bằng Airflow. Trong trường hợp thực tế, Kafka Producer sẽ liên tục lắng nghe API, vì mục đích trình diễn, chúng tôi sẽ lên lịch cho tác vụ phát trực tuyến Kafka chạy hàng ngày. Sau khi quá trình truyền phát hoàn tất, công việc Spark sẽ xử lý dữ liệu, làm cho dữ liệu sẵn sàng để ứng dụng LLM sử dụng

Tên cột	Ý nghĩa
reference_fiche (reference sheet)	Mã định danh duy nhất của sản phẩm bị thu hồi. Nó sẽ đóng vai trò là khóa chính của cơ sở dữ liệu Postgres của chúng tôi sau này.
categorie_de_produit (Product category)	Ví dụ: thực phẩm, thiết bị điện, dụng cụ, phương tiện vận tải, v.v ...
sous_categorie_de_produit (Product sub_category)	Ví dụ: chúng ta có thể có thịt, các sản phẩm từ sửa, ngũ cốc làm danh mục phụ cho danh mục thực phẩm
motif_de_rappel (Reason for recall)	Lý do thu hồi là một trong những lĩnh vực quan trọng nhất
date_de_publication	ngày sản xuất
risques_encourus_par_le_consommateur	chứa đựng những rỏi ro mà người tiêu dùng có thể gặp phải khi sử dụng sản phẩm

Clearn Data
+ Xóa cột ndeg_de_version và rappelguid
+ Kết hợp 2 cột risques_encourus_par_le_consommateur và description_complementaire_du_risque -> có cái nhìn tổng quát hơn về rủi ro của sản phẩm
+ Chia cột date_debut_fin_de_commercialisation thành 2 cột riêng biệt -> truy vấn dễ dàng hơn về thời điểm bắt đầu hoặc kết thúc quá trình tiếp thị sản phẩm
+ Xóa dấu khỏi tất cả các cột ngoại trừ link, reference numbers, and dates -> một số công cụ văn bản gặp khó khăn trong việc sử lý dấu