Detailed Email Processing Steps

1. Generate emails from JSON.
    - Reads email data from a JSON file.
    - Parses the JSON structure to extract email fields (sender, recipient, subject, body, attachments).
    - Transforms JSON data into a structured format suitable for further processing.

2. From emails and their attachments, extract the data content and push it to MongoDB.
    - Iterates through the extracted emails.
    - For each email, extracts the email body and any attached files.
    - Parses attachment content (e.g., PDFs, text files) to extract relevant data.
    - Stores the extracted email and attachment data into a MongoDB database.
    - MongoDB acts as a persistent storage for email data.

3. Run a trained LLM model to classify emails by reading the subject and body from the database
    - Retrieves email subject and body from MongoDB.
    - Feeds the subject and body to a pre-trained Large Language Model (LLM).
    - The LLM classifies the emails based on their content (e.g., 'request', 'information', 'spam').
    - Stores the classification label back into the MongoDB record for each email.

4. For records labeled as 'request', extract key fields like SSN number, request type, subtype, an
    - Filters MongoDB records to select emails classified as 'request'.
    - Applies specific extraction rules to identify key fields (SSN, request type, subtype).
    - Regular expressions or NLP techniques are used for field extraction.
    - Stores the extracted key fields as additional attributes within the MongoDB record.

5. Run a duplicate checker: first, on PII data. If it doesn't match, run it on customer email_id.
    - Retrieves PII data (e.g., SSN) from MongoDB records.
    - Compares PII data to identify potential duplicates.
    - If no duplicates are found based on PII, compares customer email IDs.
    - Sets a 'duplicate' flag in the MongoDB record indicating whether a duplicate was found.

6. For records where 'duplicate' is false, run full extraction to extract all required fields.
    - Filters MongoDB records to select emails marked as 'duplicate: false'.
    - Performs a comprehensive extraction of all remaining required fields from the email and attachments.
    - This step might involve more complex parsing and NLP techniques.
    - Stores the fully extracted data into the MongoDB record.

Generated on: 2025-03-27 09:43:56