Updated Technical Specifications with SKU Matching Using Multimodal Techniques

Project Overview (Updated)

We are building a web-based system to scrape product data from retailer sites using Browse AI, store the scraped data in a structured database, and display it on a front-end dashboard. The system will also perform multimodal content compliance and image matching checks on the scraped data using a specialized model. Additionally, we will use these multimodal techniques to assist in SKU matching against a Global SKU by cross-referencing product titles, primary and secondary images. The aim is to automate data normalization and ensure that products from various retailers are consistently matched to their corresponding Global SKU.

Project Components (Updated)

1. Web Scraping with Browse AI
2. Back-End Data Storage and Normalization with SKU Matching
3. Multimodal Content Compliance, Image Matching, and SKU Matching
4. Front-End Dashboard
5. Automation and Increased Scraping Frequency

2. Back-End Data Storage and Normalization with SKU Matching

- Global SKU and Retailer-Specific SKU Management:
- The Global SKU will serve as the central identifier across all retailers for the same product.
- Retailers often use their own SKUs, which might not match universally available identifiers like EAN or UPC. To ensure data normalization, we will perform multimodal SKU matching by comparing product titles and images (including secondary images) from the scraped data.
- Database Schema Design (with SKU Matching Considerations):
- In addition to the previously mentioned tables, ensure the following fields are in place to track and automate SKU matching:
- Products Table:
- global_sku (Primary Key for normalization)
- product_name, brand, category
- Retailer_Product_Data Table:
- retailer_sku
- product_id (Foreign Key to Products Table, Global SKU)
- title_compliance, image_match_compliance (additional compliance fields for SKU matching)
- Automation of SKU Matching:

- The system will use multimodal models to match scraped product data (titles and images) from various retailers against the Global SKU. This will automate SKU normalization, ensuring consistency in product identification across retailers.
- The multimodal model will analyze:
- Product Titles: Check for variations in product titles and determine the likelihood of a match based on keyword analysis and phrasing.
- Primary Images: Compare the primary product image to the Global SKU reference image.
- Secondary Images: Use secondary images (e.g., lifestyle images or product variant images) to confirm that the product matches the Global SKU, especially when titles may be ambiguous.

3. Multimodal Content Compliance, Image Matching, and SKU Matching

- Multimodal SKU Matching Using Titles and Images:
- In addition to the regular content compliance checks, we will leverage multimodal techniques for SKU matching by:
- Title Matching: The model will compare product titles from scraped data against known product titles associated with the Global SKU. It will check for minor variations (e.g., word order, additional descriptors) and flag any inconsistencies.
- Image Matching: The model will compare both primary and secondary images to the Global SKU’s reference images. If the images are a close match, the system will confirm that the product is associated with the correct Global SKU.
- Confidence Score for SKU Matching: The model will return a confidence score for each match, indicating how closely the retailer’s product title and images align with the Global SKU. For example:
- Title match: 0.85 (out of 1.0)
- Image match (primary): 0.90
- Image match (secondary): 0.88
- Handling Partial Matches: If a product’s title and image match to a certain threshold (e.g., 80% or higher), the system will automatically assign the Global SKU. Otherwise, it will flag the product for manual review.
- Database Storage of SKU Matching Results:
- Create or update the Retailer_Product_Data table to store SKU matching results:
- sku_match_score (float): The overall confidence score based on title and image matching.
- title_match_score, image_match_score (floats): Individual scores for the title and image matches.
- manual_review_needed (boolean): Flag for products that do not meet the required match threshold and need manual review.
- Data Workflow:

- Scraped data is passed through the multimodal model for both content compliance and SKU matching.
- The results are stored in the database alongside other product data, with the sku_match_score determining if the Global SKU can be automatically assigned.
- If the SKU match score is low, the product is flagged for manual review.

5. Automation and Increased Scraping Frequency (Updated)

- Automated SKU Matching Integration:
- After each scraping job, the scraped data (including titles and images) will be automatically submitted to the multimodal model for SKU matching.
- The SKU matching process will run as part of the compliance checks, and results will be stored in the database along with the confidence scores for each match.
- Scheduling and Batch Processing: The SKU matching process will be scheduled to run at frequent intervals (e.g., after each scraping session or during off-peak hours) to ensure that all data is normalized as quickly as possible.
- Handling Exceptions and Errors in SKU Matching:
- Set up logging and error tracking for cases where the SKU matching process fails or the confidence score is too low to assign a Global SKU.
- Alerts can be triggered for manual intervention, allowing users to manually review products that do not meet the automated matching threshold.

Summary of Key Technologies (Updated):

- PHP Laravel: Backend framework for API integration, database management, scheduling, and task automation.
- Browse AI: Scraping tool for collecting product data from retailer sites, integrated via APIs, with increased scraping frequencies to ensure data freshness.
- Multimodal Model: External service for performing content compliance, image matching, and SKU matching based on product titles and images (both primary and secondary).
- Front-End Dashboard: Real-time data display and compliance visualization, including SKU matching confidence scores and manual review flags.

Advantages of This Approach:

- Automation of Data Normalization: By using multimodal techniques to match product titles and images, the system automates the normalization of retailer-specific SKUs against the Global SKU, improving accuracy and reducing manual effort.
- Improved Matching Accuracy: Combining text-based and image-based analysis ensures that even products with slightly different titles or visual representations can be correctly matched.

- Scalability: This approach is scalable, as more products and retailers are added, and the matching logic can adapt to various types of data (e.g., new product images, title variations).