Bulk Supplier Data Import Guide

For the Replit Development Agent

Version: 1.0
Date: 2025-07-25
Author: Replit Coach Too
Status: Draft

1. Objective

This document provides a detailed technical guide for the Replit Agent to build an advanced **Bulk Supplier Data Import** feature. This tool will be used by the internal Avallen Solutions team via the Super Admin Dashboard. It will allow an admin to provide a single source (either a website URL to a product catalogue or a PDF catalogue document) and have the system automatically scrape and structure the data for both the supplier's company profile and their entire product listing.

2. Part 1: Technology & Backend Enhancements

2.1. New Technologies

- **Web Crawling:** To handle multi-page website scraping, the backend will incorporate **Scrapy**, a powerful Python framework for crawling websites and extracting structured data.
- **PDF Parsing:** To extract data from PDF documents, the backend will use **pdfplumber**, a library that excels at extracting text, tables, and other data from PDF files.

2.2. New Backend Service: BulkImportService

- **Logic:** A new, asynchronous service will be created to handle these long-running import tasks. It will be managed by our existing Celery/Redis task queue.
- **Functionality:** This service will have two primary functions: import_from_url(url) and import_from_pdf(file_path). Both functions will aim to return a single, structured JSON object containing the scraped company profile and a list of scraped products.

3. Part 2: User Interface (Super Admin Dashboard)

This feature will be added to the admin's supplier management dashboard, located at /admin/suppliers.

- **UI Component:** A new button will be added to the SupplierAdminTable view: **"Bulk Import Suppliers"**.
- **Modal Interface:** Clicking the button will open a modal with two options:

   1. **"Import from URL":** An input field for a website URL.
   2. **"Upload PDF Catalogue":** A file uploader for PDF documents.
- **Action:** The admin chooses an option, provides the source, and clicks a "Start Import" button. This will trigger the asynchronous backend service.

4. Part 3: Workflow for URL-based Bulk Import (Web Crawler)

1. **Admin Action:** The admin provides a URL to a supplier's main product catalogue page (e.g., a gallery of all their bottles).
2. **Backend Process (import_from_url):**
   - A new Celery task is initiated.
   - **Step A - Scrape Company Profile:** The Scrapy crawler first scrapes the *initial* URL for company profile information. It will look for common pages like "About Us" or "Contact Us" to find the company name, address, website, and contact email.
   - **Step B - Identify Product Links:** The crawler then analyzes the catalogue page to identify all the <a> (hyperlink) tags that appear to lead to individual product detail pages.
   - **Step C - Crawl & Scrape Each Product:** The crawler will then visit each of the identified product links one by one. On each individual product page, it will run the single-page scraping logic (using BeautifulSoup as planned previously) to extract the detailed product attributes (weight, material, dimensions, etc.).
   - **Step D - Consolidate Data:** The service consolidates all the scraped information into a single JSON object.
3. **Data Review (Human-in-the-Loop):**
   - Once the crawl is complete, the admin is notified.
   - A new view appears in the admin dashboard, presenting the scraped data in an editable form. The company profile is at the top, followed by a table of all the products that were found.
   - The admin must **review, correct, and confirm** all the data. This is a critical step to ensure data quality.
   - Upon confirmation, the backend saves the data, creating a new verified_suppliers entry and multiple supplier_products entries.

5. Part 4: Workflow for PDF-based Bulk Import

1. **Admin Action:** The admin uploads a supplier's PDF product catalogue.
2. **Backend Process (import_from_pdf):**
   - A new Celery task is initiated.
   - The uploaded PDF is processed using pdfplumber.

- **Step A - Extract Company Profile:** The service scans the first and last few pages of the PDF for text patterns matching addresses, emails, and websites to build the company profile.
- **Step B - Extract Product Listings:** The service then iterates through the entire document, using regular expressions and pattern matching to identify distinct product entries. It will look for keywords like "Product Name:", "Weight:", "Material:", and table structures to extract the attributes for each product.
- **Step C - Consolidate Data:** All extracted information is consolidated into the same standard JSON format as the URL import.

3. **Data Review (Human-in-the-Loop):**
   - The process is identical to the URL import. The admin is presented with the extracted data in an editable format.
   - The admin must review, correct, and confirm all data before it is saved to the database.

This bulk import feature will dramatically accelerate the process of building your supplier network, providing a significant operational advantage for your team.