需求規格書

# 1. 專案簡介

- 專案名稱：購物網站爬蟲排程系統  
- 版本編號：1.0  
- 撰寫日期：2025/05/26  
- 撰寫者：Alex  
 - 審核者：遠東方  
- 適用對象：開發人員、專案管理者、維運人員

# 2. 背景與目標

- 背景說明：隨著電商資料分析需求提升，需建立自動化爬蟲系統來每日擷取各大購物網站資訊。  
- 開發目的：透過 Python 搭配 Airflow 建立可排程之多網站爬蟲系統。  
- 範圍說明：本系統涵蓋家樂福、美廉社、全聯 PXMart、PXGo、Elife、愛買 RTMart 資料擷取，並將結果以檔案儲存；不包含資料分析。

# 3. 系統概觀

- 系統架構圖（可附圖）  
- 角色與使用者類型：開發者、排程維運者  
- 主要模組與功能列表：  
 - 爬蟲模組（靜態、動態、API 三類型）  
 - Airflow 排程模組  
 - 檔案儲存模組

# 4. 功能性需求（Functional Requirements）

## 4.1 爬蟲類型處理

- 識別碼：FR-001  
- 功能描述：依網站類型使用靜態、Playwright 或 API 方式取得資料  
- 家樂福：靜態爬取  
- 美廉社：動態爬取（Playwright）  
- PXMart、PXGo、Elife、RTMart：API 呼叫

## 4.2資源獨立執行

- 識別碼：FR-002  
- 功能描述：可同時啟動多個爬蟲且互不干擾（多進程或任務隔離）  
- 前置條件：Docker 中的 Airflow 正常啟動  
- 輸入：排程觸發  
- 輸出：各自的檔案結果

## 4.3啟動爬蟲

- 識別碼：FR-003  
- 功能描述：可從 Airflow DAG 排程觸發各網站之爬蟲模組。  
- 前置條件：Airflow 正常啟動  
- 輸入：DAG 排程設定或手動觸發  
- 輸出：實體檔案（EXCEL/CSV）  
- 例外處理：網站連線失敗、自動重試或記錄錯誤

# 5. 非功能性需求（Non-Functional Requirements）

- 效能需求：每日完成所有網站爬取，資源使用最小化  
- 可用性：Airflow 運作穩定  
- 擴充性：可加入新網站模組  
- 安全性：內部使用不涉登入資訊  
- 可維護性：模組分離、具備記錄與除錯機制

# 6. 使用者介面規格（UI）

- 畫面流程圖：無需圖形介面  
- 使用介面：使用 Airflow 內建 Web UI 瀏覽 DAG 狀態與手動啟動  
- UI元件需求：無需額外開發 UI

# 7. 外部介面需求（API/系統整合）

- API清單與格式：使用 PXMart、PXGo、Elife、RTMart 官方 API（依網站回傳格式整合）  
- Playwright 整合：支援啟動無頭瀏覽器動態爬蟲  
- Docker 環境：Airflow 裝於 Docker 中執行，並由 volume 儲存檔案

# 8. 限制與假設條件

- 作業系統：Windows VM  
- 瀏覽器支援：Playwright 需使用 Chromium  
- 開發語言：Python 3.12  
- 系統假設：網站結構在每日爬取間不會大幅更動，且網站無臨時更新情況

# 9. 測試規劃與驗收準則

- 基本驗收條件：  
 - 各網站之爬蟲可以最小資源運作並每日完成一次擷取  
 - 程式碼可手動啟動，也可由 Airflow 自動啟動  
- 測試方式：手動測試每個 DAG、驗證儲存檔案內容與格式

# 10. 版本控制紀錄

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 日期 | 修改者 | 修改內容 |
| 1.0 | 2025/05/26 | Alex | 初始化規格需求書 |