Gumtree Scraping Automation (n8n + Scrapfly)

1. Project Overview

Build an automated **n8n workflow** that scrapes **Gumtree NSW Farming & Veterinary job listings** using **Scrapfly**, including **phone numbers behind a login wall**, and stores results in **Google Sheets**.

- Trigger type: **Manual only**
- Storage: **Google Sheets (append-only)**
- Duplicate handling: **New-only scraping**

Target URL:

```
https://www.gumtree.com.au/s-farming-veterinary/nsw/c2121013008839
```

---

2. Technology Stack (Mandatory)

- n8n (workflow automation)
- Scrapfly API (authenticated scraping)
- Google Sheets

❌Selenium / browser automation is NOT allowed.

---

3. Authentication & Login-Wall Handling (Critical)

Phone numbers are hidden behind a Gumtree login wall.

Required approach: - Use **Scrapfly authenticated session** - Enable **session persistence (cookies reuse)** - Login handled entirely inside Scrapfly - Reuse the same session for job detail pages

Security Rules

- No credentials hardcoded in code or nodes
- Credentials must be stored ONLY in:
- Scrapfly session config, or
- n8n encrypted Credentials Manager

---

4. Data Fields to Extract

Each job listing must produce the following fields:

Field	Description
job_id	Unique identifier (from Gumtree or derived from URL)
title	Job title
url	Full Gumtree job URL
location	Job location
phone	Phone number (login-protected, if available)
creationDate	Posting date
description	Full job description
categoryName	Gumtree category

If `phone` is unavailable, leave blank or skip gracefully.

---

5. Google Sheets Structure

Sheet 1: `jobs`

- Stores full job data
- Append-only (never update existing rows)

Sheet 2: `index`

- Single column: `job_id`
- Used for duplicate detection

---

6. Incremental (New-Only) Scraping Logic

Each manual run must scrape **only new jobs**.

Process: 1. Scrape listing pages 2. Extract `job_id` 3. Check `job_id` in `index` sheet - Exists → skip - Not exists → scrape details + append

This prevents duplicates even if Gumtree reorders listings.

---

7. Pagination Requirements

- Scrape **all available pages**
- Stop when no next page is found

---

8. Recommended n8n Workflow Structure

1. Manual Trigger
2. Scrapfly request – listing pages
3. Pagination loop
4. Extract job URLs + job_id
5. Google Sheets lookup (`index`)
6. Conditional (new-only)
7. Scrapfly request – job detail page (authenticated)
8. Extract full data + phone
9. Append to Google Sheets (`jobs` + `index`)
10. Logging

---

9. Error Handling & Stability

- Retry failed Scrapfly requests
- Skip broken listings without crashing workflow
- Log:
- Total listings scraped
- Total new jobs added
- Total skipped
- Errors encountered

---

10. SQA (Software Quality Assurance) Testing

10.1 Test Environment

- n8n test instance
- Scrapfly test quota
- Google Sheets test file

10.2 Test Cases

TC-01: Manual Trigger Execution

Steps: - Run workflow manually

Expected Result: - Workflow starts without error - No automatic scheduling triggered

---

TC-02: Pagination Coverage

Steps: - Run workflow on a category with multiple pages

Expected Result: - All pages scraped - No page skipped

TC-03: New-Only Logic (Duplicate Prevention)

Steps: 1. Run workflow first time 2. Run workflow second time without new jobs

Expected Result: - First run appends jobs - Second run appends **zero** new rows

---

TC-04: Append-Only Validation

Steps: - Re-run workflow

Expected Result: - Existing rows remain unchanged - Only new rows appended

---

TC-05: Phone Number Extraction (Login Wall)

Steps: - Scrape job with phone number hidden behind login

Expected Result: - Phone number extracted correctly - No authentication errors

---

TC-06: Missing Phone Handling

Steps: - Scrape job without phone number

Expected Result: - Job is still appended - Phone field blank or null

---

TC-07: Error Resilience

Steps: - Simulate network/API failure

Expected Result: - Workflow retries or skips gracefully - Workflow does not crash

---

10.3 Acceptance Criteria

✔️Manual trigger works ✔️Only new jobs are appended ✔️Phone numbers behind login are captured ✔️ No duplicate rows ✔️No credentials exposed ✔️Logs available for review

---

11. Final Deliverables Checklist

- Importable n8n workflow JSON
- Verified Google Sheets output
- SQA tests passed
- Short usage explanation

Document Owner: Engineering Team **Prepared for:** Junior Engineer Implementation