📘 Project Guide: Smart Invoice Automation using n8n & AI

This guide helps beginners build a complete automation that extracts data from scanned and digital invoice PDFs and logs the results into a Google Sheet using n8n.

✅ Overview

We'll automate the following:

- Download invoice PDFs from Google Drive
- Extract text using native PDF tools or OCR for scanned files
- Use AI to structure the data
- Append the final output to Google Sheets

🔧 Tools Required

- **n8n Cloud or Local Setup**
- **Google Drive** (for source PDFs)
- **OCR.Space API** (for scanned PDFs)
- **Google Sheets** (for storing structured data)
- Optional: **ChatGPT/Gemini** for schema-based extraction

⍰ Step-by-Step Workflow Instructions

🔁 Step 1: Manual Trigger

- Node: Manual Trigger
- Purpose: Initiates the workflow (can later be changed to scheduled)

📂 Step 2: Download PDF File from Google Drive

- Node: Google Drive
- Operation: Download File
- Settings:
  - **Resource**: File
  - **File**: Select from list (e.g., INV-1001.pdf)
  - **Output**: Binary data

📄 Step 3: Extract Text from PDF

- Node: Extract From File
- Operation: Extract From PDF
- Input Binary Field: data
- Output: Text fields including:
  - text
  - numpages, info, etc.

🔀 Step 4: IF Node – Check if Text is Empty

- Node: IF
- Condition:

{{ $json.text }}

is empty

- Output:
    - **False** → Pass to Step 5B (Text is extractable)
    - **True** → Pass to Step 5A (Scanned image, needs OCR)

🌐 Step 5A: OCR Space API – For Scanned PDFs

(Executed **only if text is empty**)

- Node: HTTP Request
- Method: POST
- URL: https://api.ocr.space/parse/image
- Headers:
    - apikey: YOUR_API_KEY
- Body (Form Urlencoded):
    - base64Image:

data:{{ $('Download file').item.binary.data.mimeType }};base64,{{ $('Download file').item.binary.data.data }}

    - language: eng
    - isOverlayRequired: false
- Output: JSON with ParsedResults[0].ParsedText

🧠 Step 5B: Information Extractor – AI-Based Parsing

(Executed **if text is available** either from Step 3 or Step 5A)

- Node: AI Agent / Information Extractor
- Text input:

{{ $json.text }} || {{ $json.ParsedResults[0].ParsedText }}

- Schema Type: Generate from JSON Example
- Sample JSON (partial):


```
{
  "invoice_number": "INV-1003",
  "invoice_date": "2025-06-07",
  "bill_to": {
    "name": "Orbit",
    "company": "Electronics"
  },
  "items": [
    {
      "description": "Gadget B",
      "quantity": 8,
      "unit_price": 34.5,
      "total": 276
    }
  ],
  "total_due": 693.73
}
```

📊 Step 6: Append Data to Google Sheets

- Node: Google Sheets
- Operation: Append Row
- Document: Select your target Google Sheet
- Sheet: Sheet1 (or as named)
- Column Mapping:
    - Invoice Number → {{ $json.output.invoice_number }}
    - Invoice Date → {{ $json.output.invoice_date }}
    - Customer Name → {{ $json.output.bill_to.name }}
    - Description → {{ $json.output.items[0].description }}
    - Quantity → {{ $json.output.items[0].quantity }}
    - Unit Price → {{ $json.output.items[0].unit_price }}
    - Total Price → {{ $json.output.items[0].total }}

🧩 Automation Flow Recap

Manual Trigger

↓

Download PDF (Google Drive)

↓

Extract PDF Text

↓

IF text is empty?

↙ ↘

OCR API Info Extractor

↓ ↓

--------> Append to Google Sheets

🧠 Prompts for AI Extractor (Helpful Tip)

Prompt Example to generate JSON Schema in ChatGPT:

“Extract structured data from this invoice text into a JSON format containing fields like invoice number, date, customer name, product line items, and totals.”

📌 Tips for Beginners

- Always test with both **scanned** and **text-based** PDFs.
- Use **base64Image field** only when sending OCR request.
- Keep column names in Google Sheets **identical** to mapped values.
- Use Manual Trigger for initial testing; later convert to **Cron (Scheduled Trigger)**.