**Invoice data extraction with LlamaParse and OpenAI**

**Workflow: Intelligent Invoice Data Extraction & Reconciliation**

**Purpose:**  
This workflow automates the extraction of structured data from invoice PDFs received via email. It uses advanced AI and parsing tools to convert unstructured PDF content into structured data and then updates a reconciliation spreadsheet.

**Key Components & Steps**

1. Email Monitoring & Triggering

* **Receiving Invoices (Gmail Trigger):**
  + **Purpose:** Monitors Gmail for incoming emails that contain PDF attachments (invoices) and have not been processed (i.e., missing the "invoice synced" label).
  + **Configuration:** Uses filters (e.g., sender, attachment type) to ensure only relevant emails trigger the workflow.
* **Wait to stay within service limits:**
  + **Purpose:** Introduces a delay to ensure API rate limits are not exceeded when processing invoice data.

2. PDF Download and Extraction

* **Download & Extract:**
  + **Download Invoice PDF:** The email attachment is downloaded.
  + **Extract from File:** The downloaded PDF is processed to extract its text content using a PDF-to-text converter.

3. Advanced PDF Parsing with LlamaParse

* **Upload to LlamaParse:**
  + **Purpose:** Sends the PDF to LlamaParse (a cloud service) which is specialized in handling complex PDFs (including tables, figures, and embedded objects).
  + **Operation:** Uses a multipart-form-data HTTP request to upload the file.
* **Get Processing Status:**
  + **Purpose:** Polls the LlamaParse API to check if the PDF has been processed and is ready for further analysis.

4. Data Extraction and Structuring

* **Apply Data Extraction Rules (LLM Chain):**
  + **Purpose:** Uses a language model (via an AI agent) to extract structured invoice details from the markdown output of LlamaParse.
  + **Task:** Extract key fields such as invoice date, invoice number, supplier and customer details, line items, and totals.
* **Structured Output Parser:**
  + **Purpose:** Validates and formats the extracted data into a well-defined JSON schema.
  + **Schema Fields:** Invoice date, invoice number, purchase order number, supplier/customer details, line items, subtotal, and total price.
* **Map Output:**
  + **Purpose:** Consolidates the structured data for further processing or storage.

5. Updating Data in Google Sheets

* **Append to Reconciliation Sheet:**
  + **Purpose:** Appends the structured invoice data into a designated Google Sheet (Invoice Reconciliation).
  + **Integration:** Uses the Google Sheets node with proper mapping of invoice fields.

6. Labeling to Prevent Duplicate Processing

* **Add "invoice synced" Label (Gmail Node):**
  + **Purpose:** Updates the processed email by adding a label ("invoice synced") to prevent reprocessing of the same invoice.

7. Supporting and Documentation Nodes

* **Sticky Notes:**
  + **Sticky Note1:** Describes the purpose of watching for invoice emails.
  + **Sticky Note2:** Explains the role of LLMs in data extraction.
  + **Sticky Note3:** Provides insights on why built-in PDF converters may be insufficient.
  + **Sticky Note4:** Details how to add a label to avoid duplicate processing.
  + **Sticky Note5:** Summarizes the workflow process and provides links for additional documentation or support.

**Data Flow Diagram (High-Level)**

1. **Gmail Trigger (Receiving Invoices) →**  
   **Download & Extract PDF → Upload to LlamaParse → Get Processing Status → Apply Data Extraction Rules**  
   **→ Structured Output Parser → Map Output → Append to Reconciliation Sheet**
2. **Processed Email → Add "invoice synced" Label**
3. **Monitoring and Delay (Wait node) ensures API limits are respected.**

**Benefits and Key Takeaways**

* **Automation & Accuracy:**  
  Automatically extracts invoice data from PDFs, reducing manual errors and streamlining data entry.
* **Advanced Parsing:**  
  Leverages LlamaParse for detailed PDF parsing, ensuring even complex documents with tables and figures are processed accurately.
* **Structured Data:**  
  Uses an AI agent to extract key invoice fields into a structured JSON format, which is then easily imported into spreadsheets or databases.
* **Efficiency:**  
  By labeling processed emails, the workflow avoids duplicate processing, ensuring efficient operations.
* **Documentation:**  
  Embedded sticky notes provide clarity and guidance on workflow configuration and customization.