**Use case description**: The objective of this Use case is to process unstructured Invoice Documents. This usecase contains classifying the documents to Invoices and non-invoices and Get structured, query-ready data using Databricks and AI-based document processing.

This structured data is intended for downstream reporting, analytics, and financial reconciliation.

**Business Problem**

Organizations receive large volumes of PDF documents (invoices, receipts, contracts, and other files) in various formats which makes their work very difficult.

Traditional Document processing has few challenges:

1. Difficulty in extracting data from complex invoice formats
2. Validation is often ignored
3. Overhead building + maintaining efficient ETL

Our solution include:

1. Identifying which files are invoices vs non-invoices.
2. Extracting structured fields (Invoice Number, Customer Name, Invoice Issuer Name, Invoice Data, Description of the product or service, Number of Units, Currency and Total Amount of the Invoice) from invoices.
3. Handling multi-language documents.
4. Processing documents incrementally as they arrive in the volume
5. Ensuring accuracy through validation

**Personas Includes:**

1. Finance & Accounting Teams - They spend significant time manually verifying, reconciling, and reporting on invoices. This use case helps them with Instant access to accurate, structured invoice data speeds up month-end closing, reduces manual errors, and improves compliance readiness.
2. Operations Managers - They oversee efficiency and cost control in document processing workflows. This use case benefits them by lower processing costs, faster turnaround times, and measurable productivity gains.
3. Compliance & Audit Teams - They need traceable, accurate records for regulatory requirements and internal audits. This use case provides a central, trustworthy repository of invoice data with full audit trails which is helpful for them

**Goals & objectives**

* Scanning the document to get the text from the scanned pdf documents using ai\_parse\_document function
* Classify PDFs into invoices and non-invoices using ai\_classify function
* For invoices extract structured fields using Information Extraction Agent
* Enable multi-lingual support for global invoice processing using ai\_translate function
* Automate the whole pipeline in the databricks jobs

**High-level Architecture**

**Data Flow**

1. File Ingestion

* PDFs stored in Databricks Volumes
* Databricks Job triggered on file arrival

1. Document Scanning

* Using ai\_parse\_document to extract raw text from the scanned pdfs

1. Classifying the Text

* Using ai\_classify function to classify the text and add a flag to the text whether it is invoice or non-invoice

1. Information Extraction

* Extract structured data from the invoices and leave the non-invoices by using the flag. Using Information Extraction Agent in databricks to get the structured data from the invoices

1. Validating the extracted data

* Validating the extracted data with the knowledge base that is provided. Check if the extracted entity is present in the knowledge base and give the nearest match for the not present entities

1. Orchestration

* Use Databricks Workflows to automate the process using file arrival

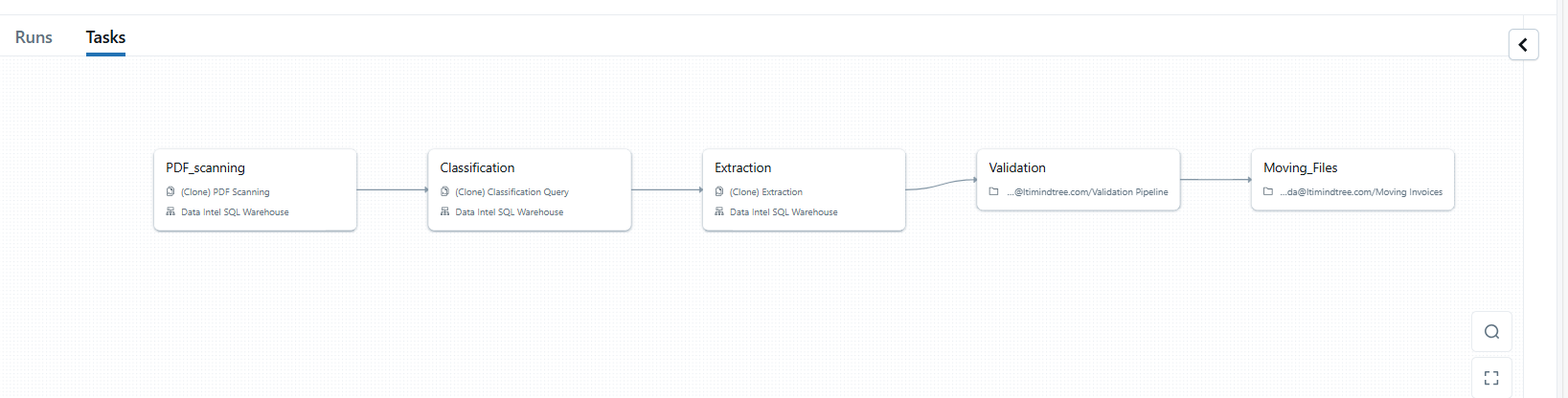
**Databricks Workflow**

**A diagram of data flow

AI-generated content may be incorrect.**

**A diagram of data flow

AI-generated content may be incorrect.**

**Lakeflow Job**

**PDF Scanning**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Classification**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Extraction**

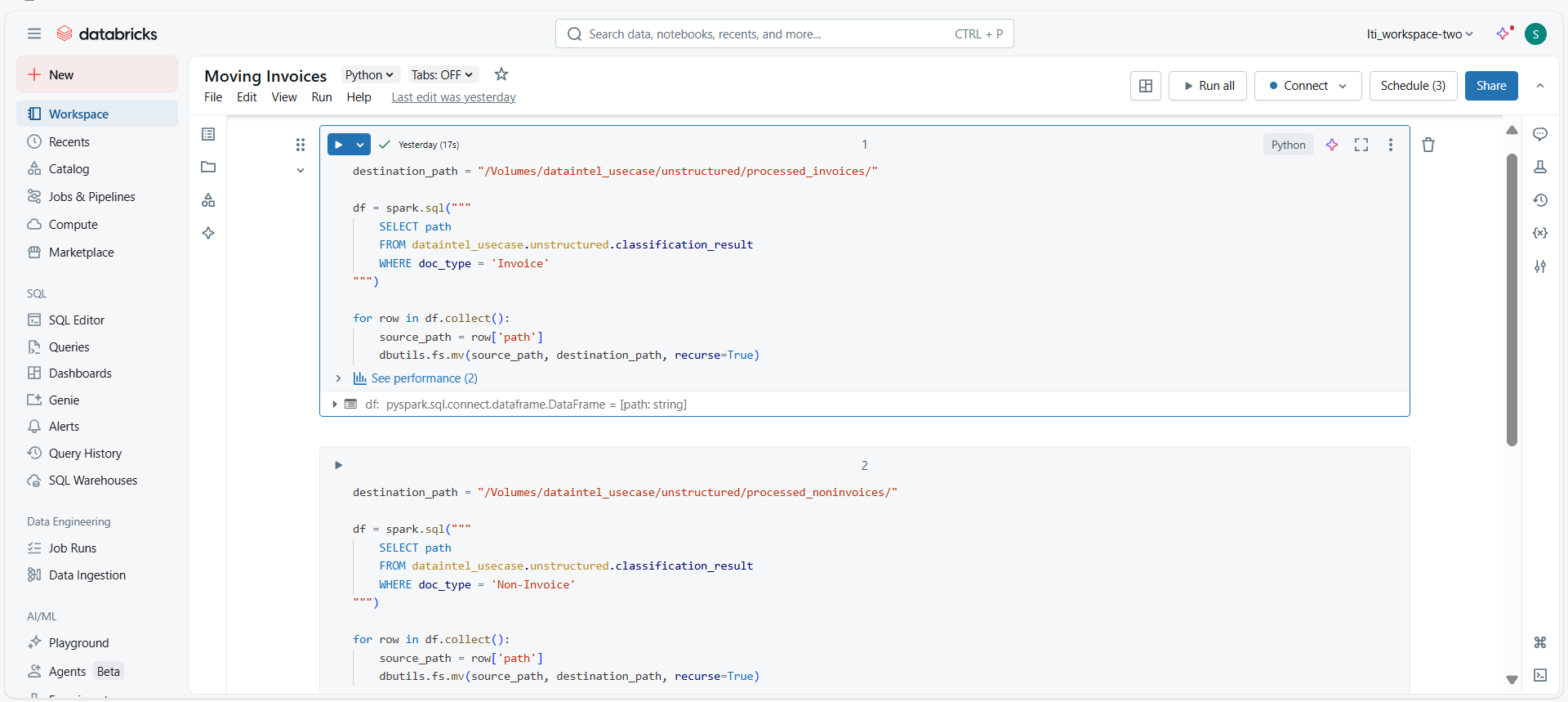
**A screenshot of a computer

AI-generated content may be incorrect.**

**Validation**

****

**Moving Invoices and Non-Invocies to separate folders**

****