

# Mydigibill: Receipt and Invoice Digitizer

# Our Team



**Mayank Kumar**  
Team Lead



**Pranitha  
Piduru**  
System  
Architect



**Tisha Gurjar**  
Frontend Dev.



**Goutham  
Mourya**  
Backend  
Developer



**Viabhavi D.**  
Presentation  
Designer



**Niharika  
Rangu**  
Technical  
Writer

# Problem Statement

Build a AI driven system that automatically scans, extracts, and digitizes information from receipts and invoices using :

- OCR (Optical Character Recognition)
- NLP-based field extraction.

The digitized data is stored in a structured format, making it easy to search, analyze, and integrate with accounting or ERP systems.



# Objectives & Challenges

Objectives	Challenges
1. Prone to Loss, errors, and manual entry delays.	----
2. OCR Accuracy: $\geq 85\%$ correct text extraction on test receipts	Noise in Receipts/Invoices. Poor Image Quality, Damaged Text inputs.
3. Field Extraction: $\geq 90\%$ accuracy on date, vendor, and total fields	Variation of Templates, missing subtotal, Different date formats
4. Validation: Detects duplicates with $\geq 95\%$ reliability.	----

# Proposed Solution

**Mydigibill** automates the digitization and processing of physical and digital receipts and invoices using OCR, re and NLP, reducing manual effort and improving accuracy in financial operations.

***Data Ingestion and  
Preprocessing***

***Data Extraction and  
Validation***

***Database Storage***

***Dashboard & Analytics***

***Integration Module***

# Additional Features

- Secure login & authentication system
- Multi-language support for accessibility.
- Chat with the database

● Language

GB English

▼

GB English

IN हिंदी (Hindi)

IN தமிழ் (Tamil)

IN తెలుగు (Telugu)

IN বাংলা (Bengali)

IN मराठी (Marathi)

Sign in with Google

OR

Full Name

John Doe

Email Address

your.email@example.com

Password

\*\*\*\*\*

👁

Confirm Password

\*\*\*\*\*

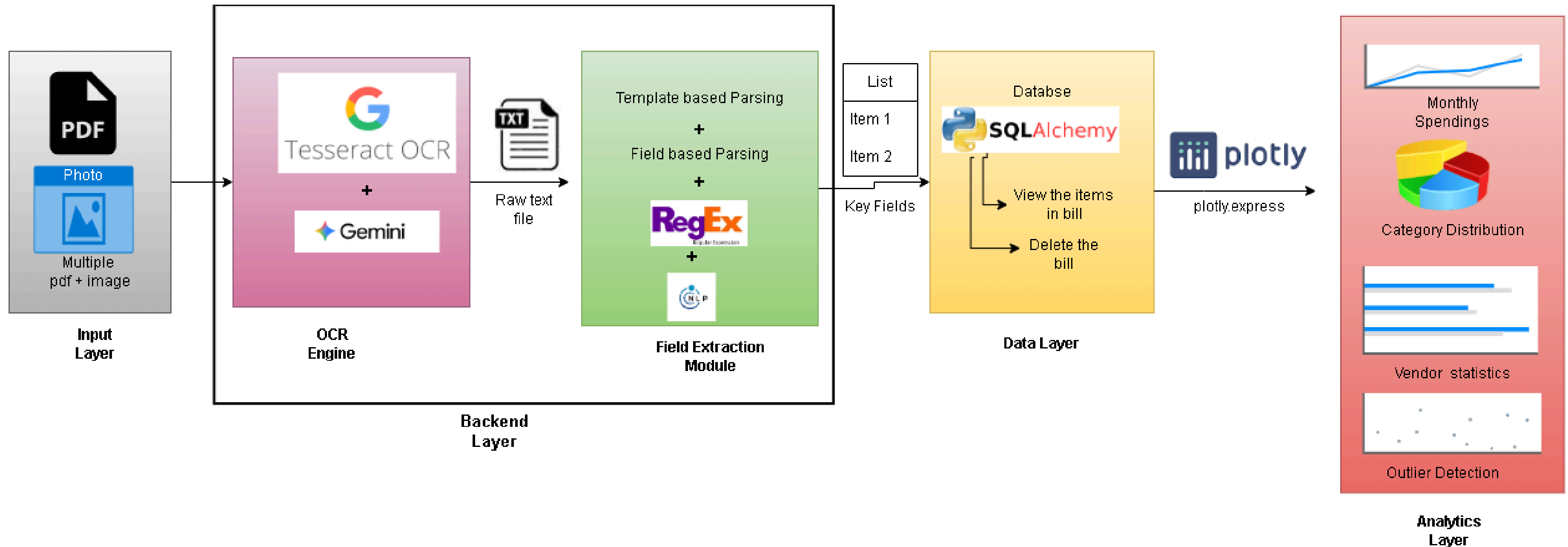
👁

Sign Up

Already have an account? [Login](#)

Login

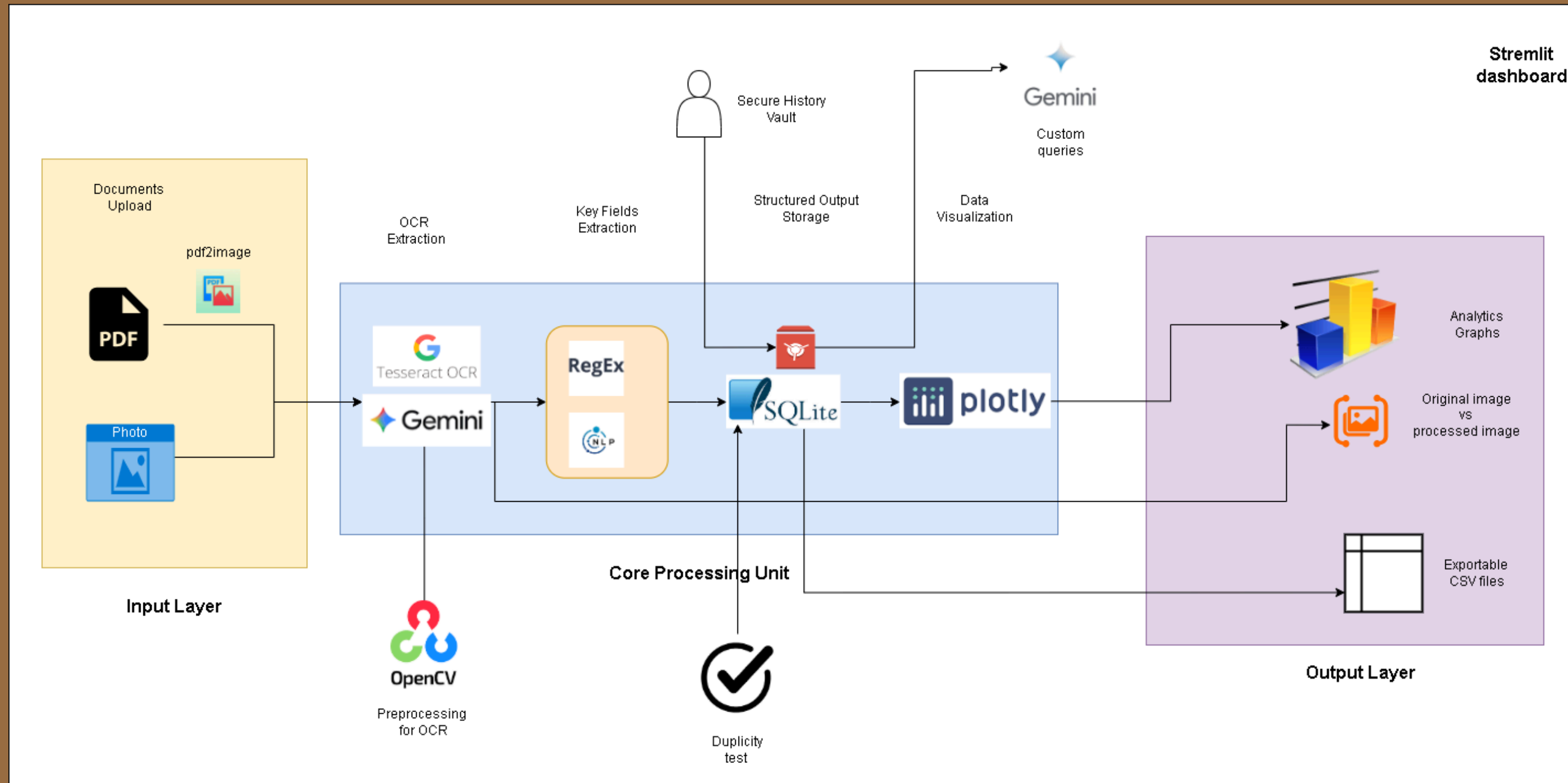
# System Workflow





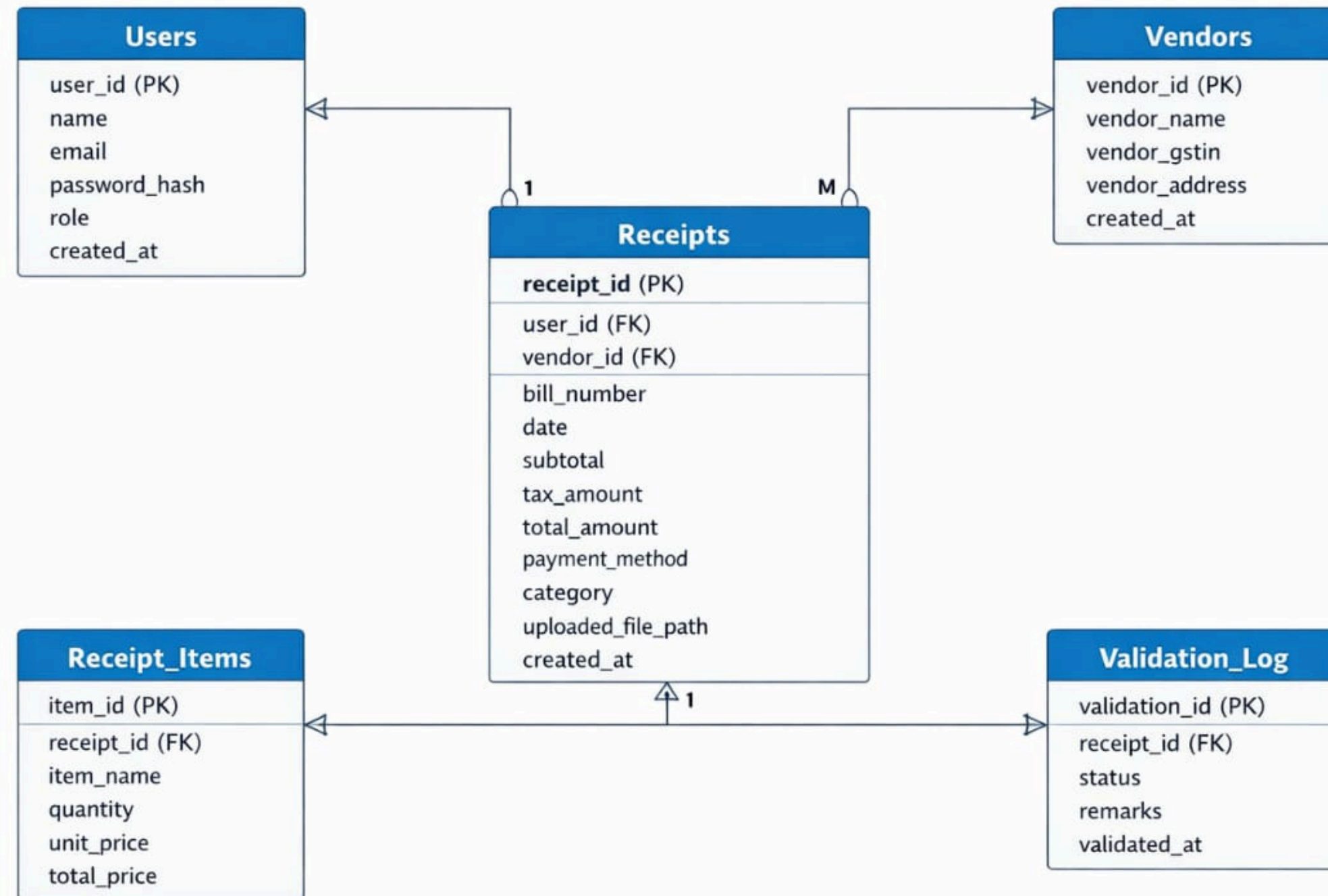
# Architecture Diagram

## Modular Diagram





# Schema Diagram



# Document Ingestion & OCR

- Supports multiple receipts upload (PDF, image)
- Detects duplicates via hash(**hashlib**).
- Automated OCR-based text extraction from scanned receipts and invoices.
- Uses both Pytesseract + Gemini to extract raw text.



## Upload Receipt

Upload receipt image or PDF



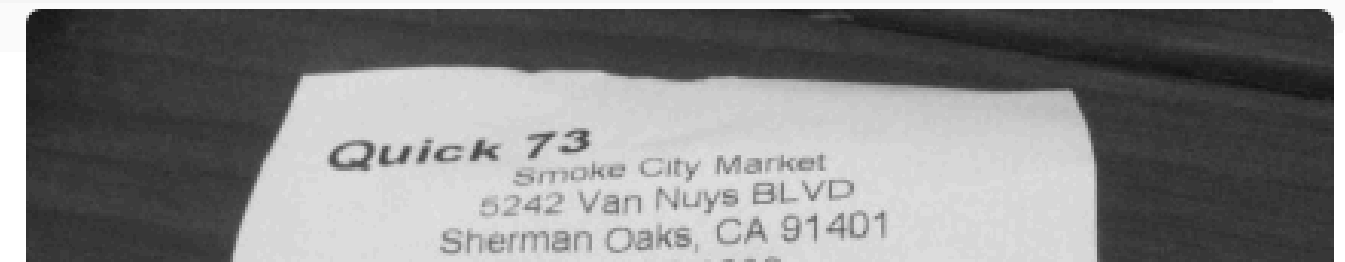
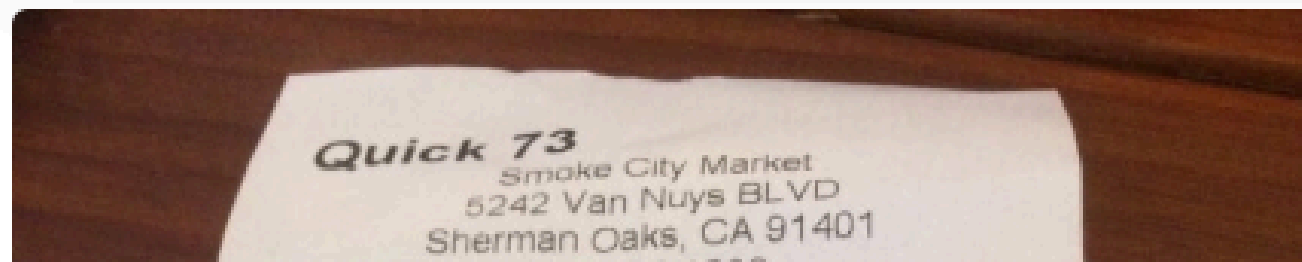
Drag and drop file here

Limit 200MB per file • PNG, JPG, JPEG, PDF

Browse files



receipt13.jpg 88.7KB



# Field Extraction & Validation

- NLP/Regex-based extraction for date, vendor, invoice ID, tax, total amount, line items.
- Field-level parsing for date, vendor, amount, tax, and line items.
- Validate extracted totals (subtotal + tax = total).
- Capable of Handling multiple layouts. (template based parsing)

## Receipt Summary


	Bill ID	Vendor	Category	Date	Subtotal (₹)	Tax (₹)	Amount (₹)
0	234673 #1	Smoke City Market	Food	2015-06-26	59.11	5.32	64.43



## Item-wise Details

Item	Price
Beef Ribs	36.86
Pulled Pork Sand	8.5
*Reg	0
Creamed Corn[Small]	3.35
Longhorn Mac & Cheese[sm]	4.2
Chocolate Cake[Slice]	3.95
Fountain Soda	2.25

# Database Storage

- Digitized data is stored in a structured format in a SQL database (SQLAlchemy).
- Support search by vendor, date, or amount.
- Export to CSV/Excel.
- Support Items wise Details info
- Allows Entries deletion for better Database Management.

 Stored Receipts

 Filter Receipts 

Bill ID

Vendor

Category

Filter by ID...

Filter by vendor...

All

Subtotal (₹)


Tax (₹)


Total (₹)


Filter by subtotal...


Filter by tax...

Filter by total...

 Delete?	Bill ID	Vendor	Date	Total (₹)	Tax (₹)	Subtotal (₹)	Category
<input type="checkbox"/>	1/A-297424	Subway	2019-05-01	₹13.22	₹1.23	₹11.99	Food
<input type="checkbox"/>	2763	Loaded Cafe	2019-02-23	₹48.53	₹4.21	₹44.32	Food
<input type="checkbox"/>	118888	Taco Maria	2018-01-23	₹45.79	₹3.29	₹42.50	Food
<input type="checkbox"/>	GUEST-CHECK-20170616-	Thai Gusto Restaurant	2017-06-16	₹30.15	₹2.67	₹27.48	Food
<input type="checkbox"/>	128	Grotto Pizzeria & Tavern	2017-05-12	₹45.58	₹2.58	₹43.00	Food
<input type="checkbox"/>	234673 #1	Smoke City Market	2015-06-26	₹64.43	₹5.32	₹59.11	Food
<input type="checkbox"/>	177437	Panda Express	2013-03-16	₹164.89	₹13.89	₹151.00	Food

 Delete Selected Receipts

 Filtered Results

 Filtered Total

7

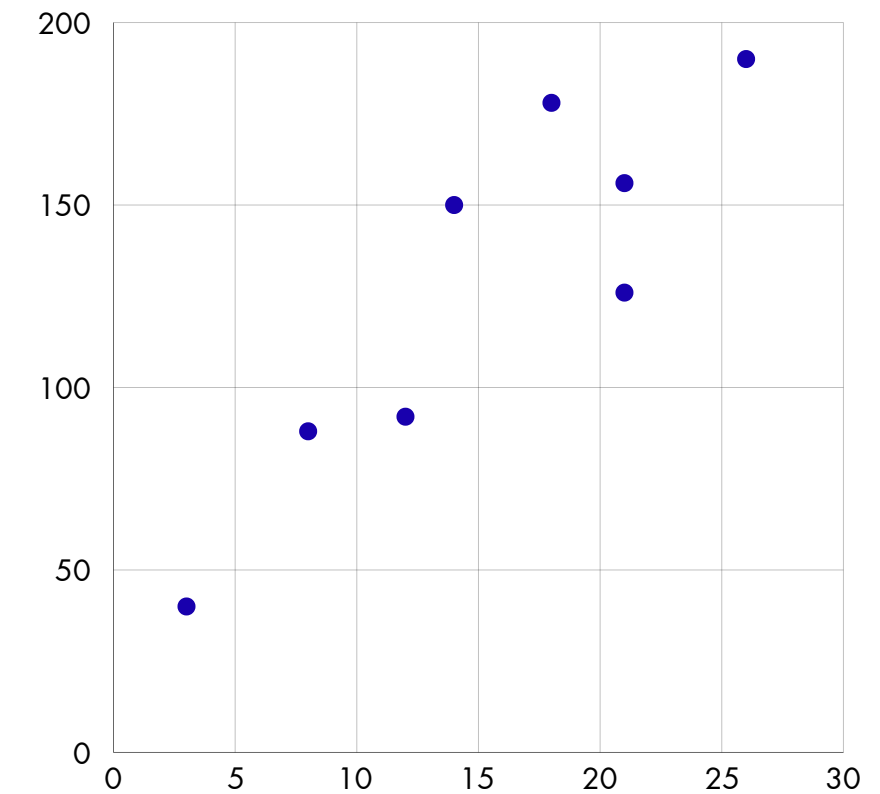
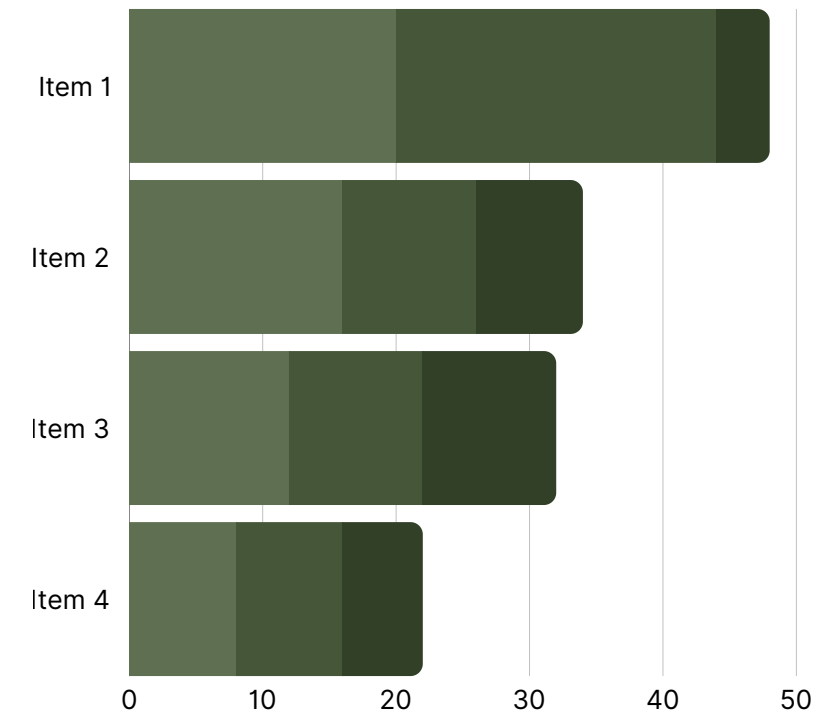
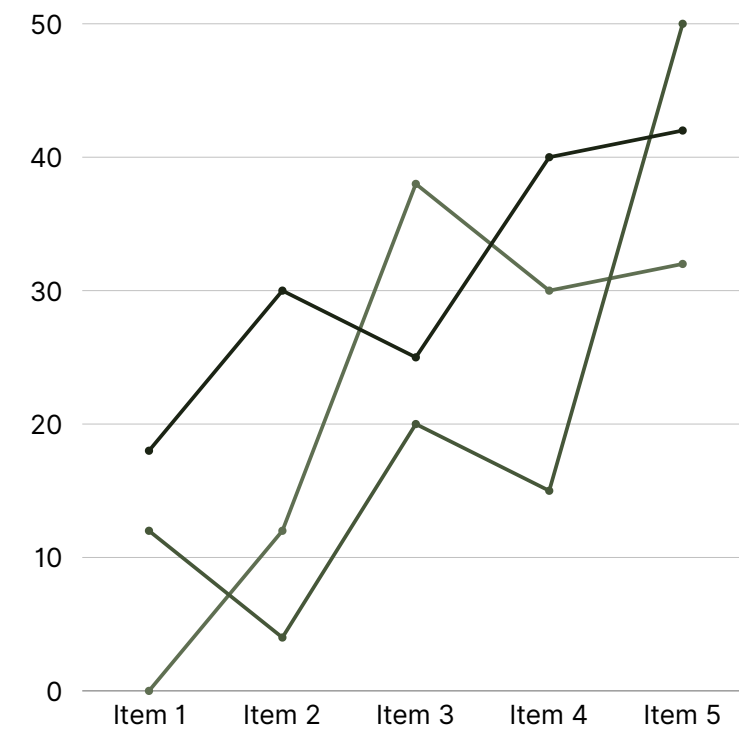
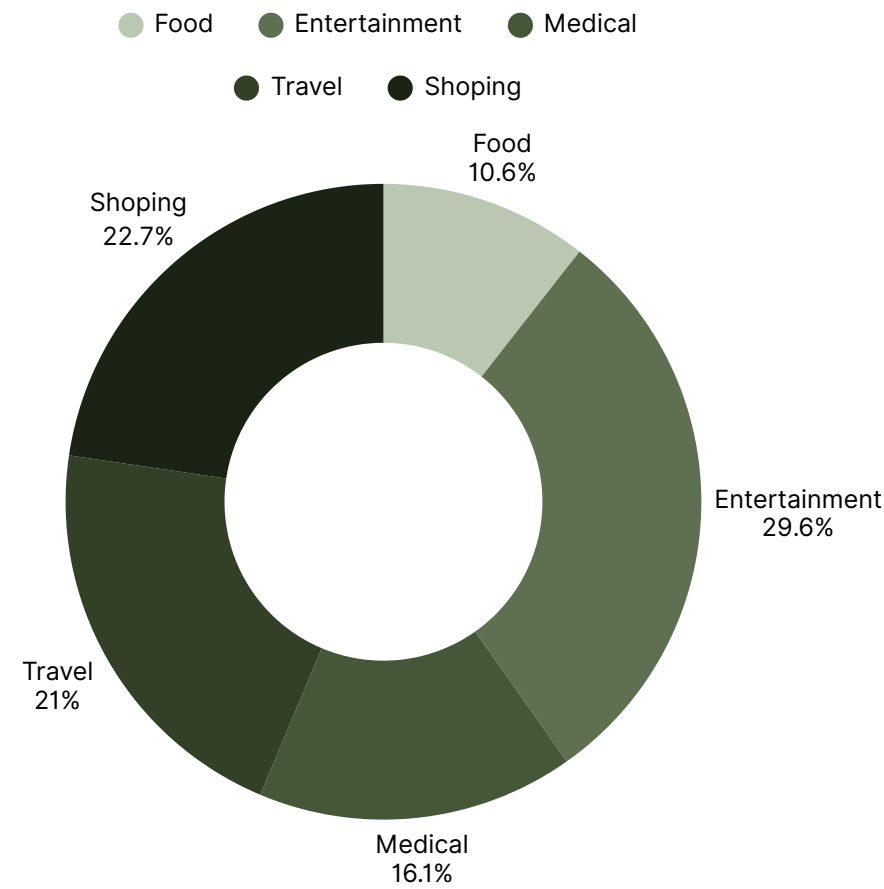
₹412.59

# Dashboard & Analytics

- Uses KPI cards to showcase the top results.
- Streamlit based Web dashbard.
- Clean, Modern and Rsponsive Interface
- Uses Multiple Charts to showcase the time depandant and other stats.
- Provides AI insights along corresponding to charts.



# Analytics



## Category Distribution

Help users understand spending distribution

## Monthly Spendings

Displays month-wise expense trends.

## Vendors Statistics

Compares spending across vendors.

## Outlier Detection

Identifies unusual expense values.

# Integration with ERP systems

- Deployment:

<https://mydigibill.streamlit.app>

- Enables communication between the digitizer and external ERP systems.
- Supports **ERPNext**, a open Source ERP system.
- API-based synchronization
- Maintains data Integrity.

## Receipt Vault Analyzer API 1.0.0 OAS 3.1

/openapi.json

REST API for ERP integration and external data access

### default ^

GET / Read Root

GET /api/v1/receipts Get Receipts

GET /api/v1/receipts/{bill\_id} Get Receipt

POST /api/v1/erp/sync Sync To Erp

POST /api/v1/ocr/process Process Image



# Technology Stack



**Frontend**  
Streamlit

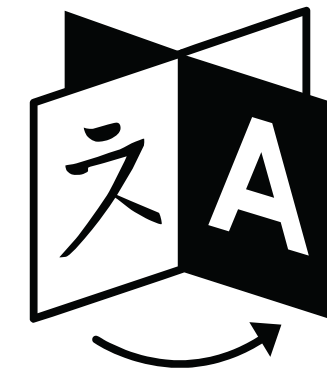


**Backend and frameworks**

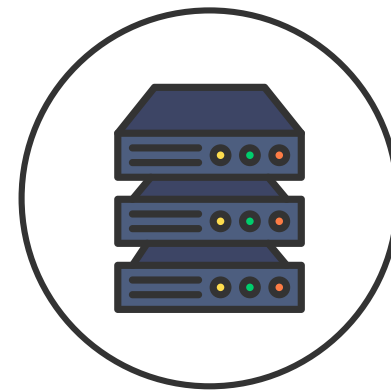
Python,  
Numpy, Pandas,  
Pytesseract, Re



**Data Vizualization**  
plotly.express



**LLM used**  
Gemini(LLM)



**Database:**  
sqlite3,  
SQLAlchemy

# References

- 8 hidden costs of outdated accounting systems
- Tesseract OCR Best Practices
- Google's Document AI (Conceptual)
- spaCy: Industrial-Strength NLP
- InvoiceNet (Academic Paper)
- Regex for Financial Data: Reliable patterns for dates, amounts, IDs
- Perceptual Hashing (ImageHash Library) for near-duplicate image detection

THANK  
YOU

