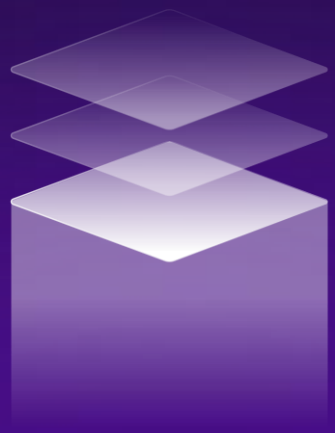
VIVIDCLOUD CASE STUDY

GenAI Enabled Tabular Data Extraction

An OCR-Driven Pipeline for SQL Based RAG Applications

Executive Summary



Organizations often face significant challenges when extracting and analyzing large amounts of tabular data embedded in PDFs, a common format for business documents. This case study presents a technology solution that leverages Optical Character Recognition (OCR) to extract tabular information from PDFs, normalizes it, and loads it into a SQL database. The solution also incorporates an SQL agent powered by natural language processing (NLP), allowing users to query the data using plain language, eliminating the need for complex SQL queries.

By automating the extraction, normalization, and querying processes, this solution enhances data accuracy, reduces manual effort, and empowers non-technical users to access and analyze data more effectively. The case study details the problem context, the implementation process, and the measurable results, illustrating how this approach can transform data handling for organizations dealing with unstructured tabular data.

Key Challenges

- **Inconsistent Data Extraction:** Manual extraction processes were inconsistent, leading to varying data quality and errors that impacted downstream processes
- **Lack of Automation**: The absence of automated tools for data extraction and normalization resulted in time-consuming and labor-intensive workflows.
- **Complex Data Querying**: Non-technical users struggle with traditional SQL queries, limiting data accessibility and insight generation.
- **Scalability Issues**: The manual processes in place were not scalable, limiting the organization's ability to efficiently handle increasing volumes of data.

VividCloud's Solution

To address these challenges, VividCloud developed a comprehensive and automated solution designed to streamline the extraction, normalization, and querying of tabular data from PDF documents. The solution is built around three key components: Optical Character Recognition (OCR) technology, data normalization processes, and a natural language-enabled SQL agent.

✓ Automated Data Extraction with OCR

VividCloud implemented advanced OCR technology capable of accurately extracting tabular data from a wide variety of PDF formats. This automation drastically reduced the manual effort required and minimized errors in the data extraction process, ensuring a higher level of data consistency and accuracy.

✓ Data Normalization and Structuring

Once extracted, the data was automatically normalized and structured into a standardized format suitable for SQL databases. The solution included sophisticated algorithms to handle various data formats and inconsistencies, enabling seamless integration across different departments and systems within the organization.

✓ Natural Language SQL Querying

To empower non-technical users, we integrated GenAI technologies such as natural language processing (NLP) and retrieval-augmented generation (RAG) to construct the SQL agent. This agent allows users to query the database using plain language, eliminating the need for complex SQL syntax and making the data more accessible to a broader range of users. This feature not only democratized data access but also sped up the decision-making process by allowing faster and more intuitive querying.

✓ Scalable and Efficient Workflow

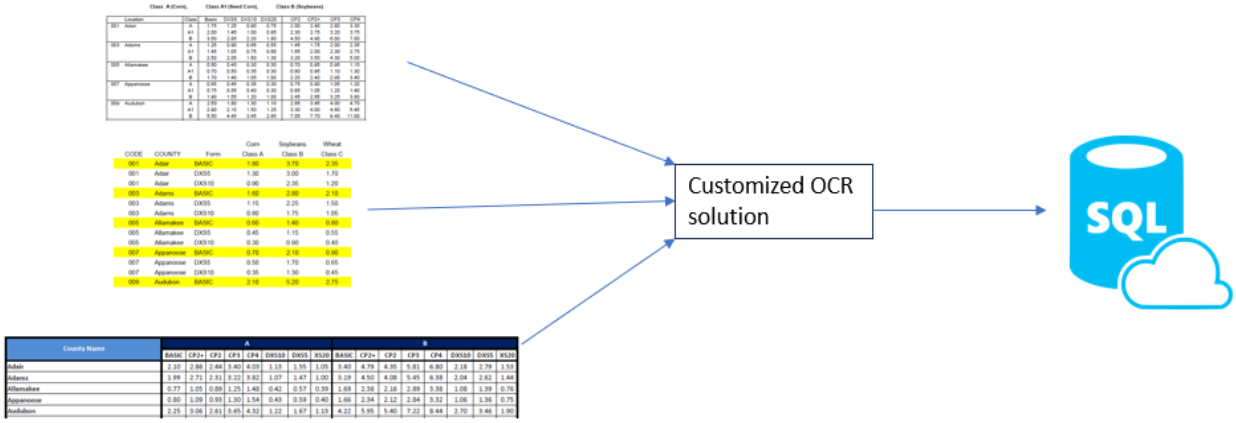
The solution was designed with scalability in mind, allowing the user to handle increasing volumes of data without sacrificing performance or accuracy. By automating key processes, the organization could reallocate resources to more strategic tasks, further enhancing operational efficiency.

Data Ingestion Pipeline:

• PDFs containing tabular information in various formats are ingested using customized OCR pipelines.

• The pipelines extract, clean and normalize the data into a unified data model.

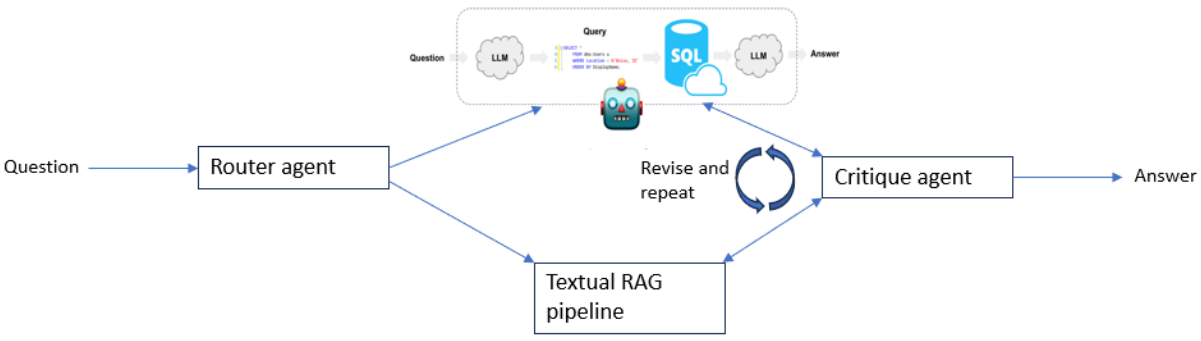
• The normalized data is then loaded into an SQL database.



SQL Agent and RAG workflows:

• Using a chat interface, users can ask questions regarding the tabular or textual data.

• An agentic system routes the question to the correct sub-system. In the case where the question is related to the tabular information, it's routed to a SQL agent. In the case where the question relates to textual data, it gets routed to a RAG system.



Results and Benefits

- **Improved Data Accuracy:** The automated OCR and normalization processes significantly improved data accuracy, reducing errors by a significant amount compared to the previous manual methods.
- **Time Savings:** The solution reduces the time required for data extraction and preparation significantly, allowing an organization to process large amounts of data in a fraction of the time previously needed.
- **Enhanced Accessibility:** The natural language SQL querying empowered a broader range of users to interact with the data, leading to an increase in the number of data driven decisions made by non-technical staff.
- **Cost Efficiency:** By reducing manual labor and improving data processing efficiency, the client achieved reduction in operational costs related to data management.

Technologies Used

- Python
- Langchain
- AWS Bedrock
- Generative AI
- Agentic based systems
- NLP
- PostgreSQL
- OCR

About VividCloud

VividCloud is a software development company focused on cloud and IoT. AWS is our cloud platform of choice, and we are an Advanced Tier APN Services Partner. We bring fully managed teams that free our client s from day to day oversight responsibilities.

VividCloud is based in Brunswick Maine, with 100% of our people onshore in the US.

Contact Us

VIVID CLOUD

Our center of gravity is in Brunswick Maine

HEADQUARTERS
150 Admiral Fitch Ave
Brunswick, ME 04011

NEW HAMPSHIRE
100 Domain Drive
Exeter, NH 03833

MASSACHUSETTS
85 Swanson St.
Boxborough, MA