Data Mining in Financial Documents

Problem

The audit project team in a road construction company needed invoice data from the last 10 years, cataloged and with the main information structured in tables, to carry out the relevant analyses for the road concession return.

The data consisted of more than 92,000 invoices.

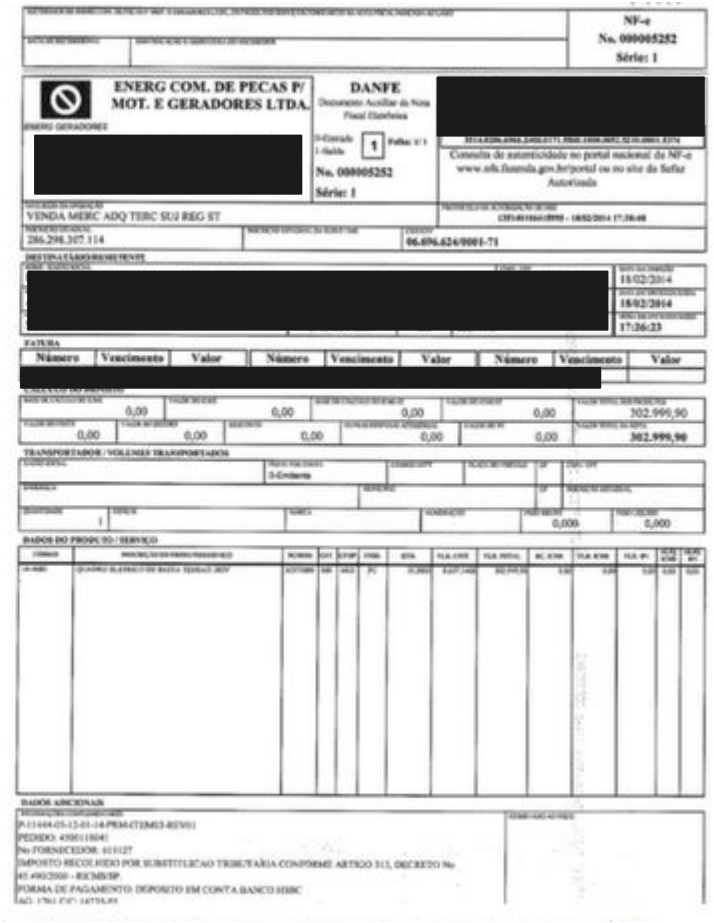
With 5 trainees working 10 hours per day, according to the manager’s estimate, it would take 48 weeks to complete the processing of all the data, at a total cost of R$ 1,306,800.00.

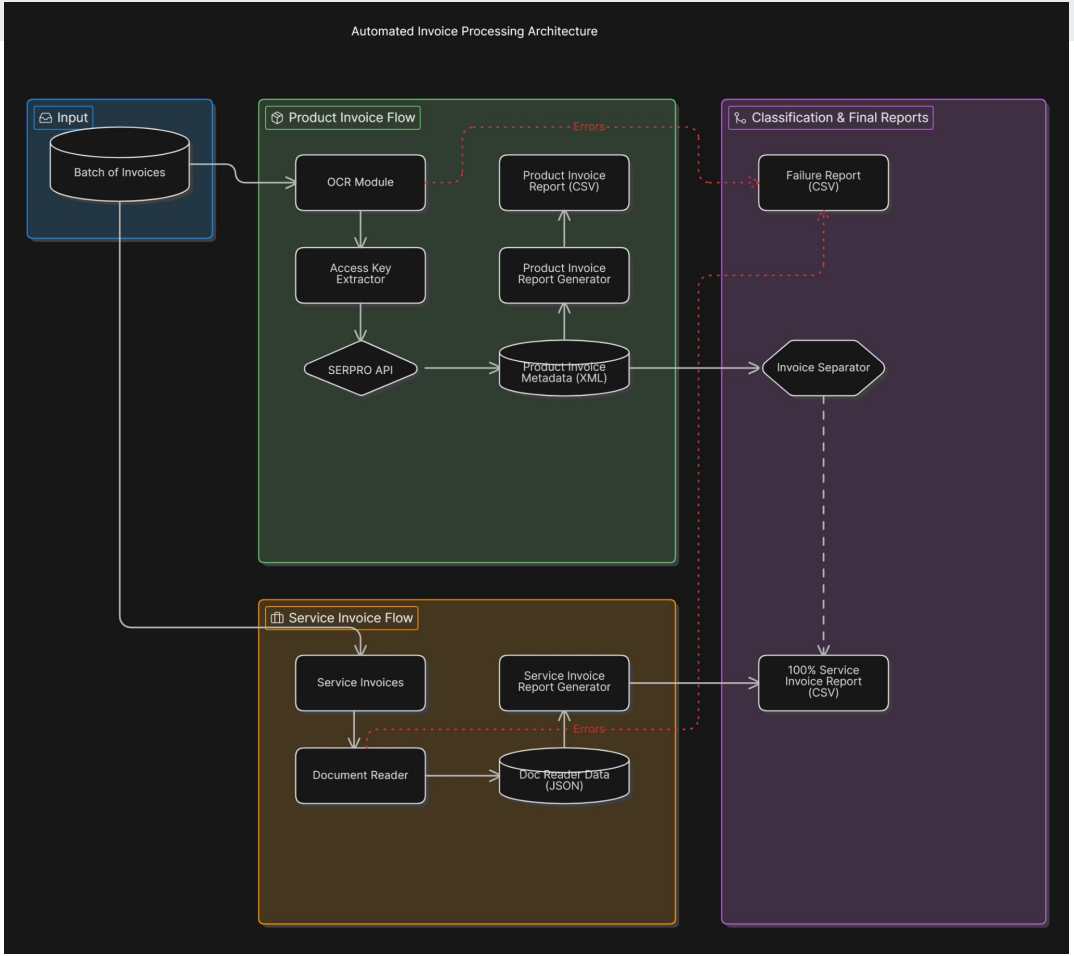
However, with this estimate, it would not be possible to meet the project delivery deadlines.

PROCESS

Data to extract:

- Issuer CNPJ
- Issuer name
- Invoice amount (value)
- Issue date
- Invoice number
- Observations/description (for product invoices)




Automated Invoice Processing Architecture
Input
Batch of Invoices
Product Invoice Flow
Errors
OCR Module
Product Invoice Report (CSV)
Access Key Extractor
Product Invoice Report Generator
SERPRO API
Product Invoice Metadata (XML)
Classification & Final Reports
Failure Report (CSV)
Invoice Separator
Service Invoice Flow
Service Invoices
Service Invoice Report Generator
100% Service Invoice Report (CSV)
Errors
Document Reader
Doc Reader Data (JSON)

Technologies Used

- Language: python
- Azure Form Recognizer OCR + Developed Structured Text Data Reconstruction
- Access key extraction is a regex module.
- Document Reader (Another project developed for data extraction from unstructured documents), built using traditional natural language processing techniques and LLM-based workflows with LangChain.
- Report Generators — Python-based system built to organize and structure the final output data according to the partner's required format.

Final Results

1. File Processing Summary

Category	Quantity
Total Files	92,964
Duplicated or Corrupted Files	4,238
Files Available for Processing	88,726

2. Invoice Classification Results

Category	Quantity
Product Invoices	30,243
Service Invoices — 100% Processed	33,740
Service Invoices — Partially Processed	21,530
Failures	3,213
Total	88,726

Final Results

3. Complete vs Manual Review Analysis

Category	Quantity	Percentage
Fully Completed Invoices	63,983	72.11%
Invoices Requiring Manual Analysis	24,743	27.89%

4. Data Extraction Performance

Category	Quantity	Percentage
Total Data Points	473,873	—
Data Successfully Extracted	408,333	86.17%
Missing Data	65,540	13.83%

5. Cost and Time Comparison

Cost Type	Cost	Time (Weeks)	Time (Hours)
Manual Processing	R$ 1,306,800.00	48	12,000
AI Development	R$ 120,000.00	3	120

Overall Conclusions

- 72.11% of invoices were fully automated.
- 27.89% still required manual analysis.
- 408,333 data fields were successfully extracted.
- Extraction success rate reached 86.17%.
- AI automation reduced processing time from 48 weeks to 3 weeks.
- Estimated operational costs were reduced by more than 90% compared to manual processing.

Document reader