Relatório de Projeto
SkyNET - I2A2
Processamento Fiscal Inteligente

Visit our page!

Este documento detalha o desenvolvimento e a arquitetura da plataforma SkyNET-I2A2, uma solução de Agentes de IA concebida para endereçar os desafios críticos da gestão fiscal no Brasil, transformando um centro de custo operacional em um ativo de inteligência estratégica.

Grupo: SkyNET

Integrantes do Grupo (em ordem alfabética)**:**

- Alberto Côrtes - cortes.albert06@gmail.com
- Ernani Fantinatti (Líder do Grupo) - ernanif@fantinatti.com
- Fábio Rhein - fabiorhein@gmail.com
- Luiz Garcia - duduluiz23@gmail.com

1. Descrição do Tema Escolhido (O que foi feito)

O grupo desenvolveu uma plataforma *end-to-end* que automatiza o ciclo completo de processamento de documentos fiscais brasileiros. A solução não apenas executa tarefas, mas orquestra um fluxo de trabalho financeiro-fiscal completo:

- **Ingestão e Captura (O "Novo Malote Digital"):** Recebimento e digitalização de documentos em múltiplos formatos (XML, PDF e imagens), eliminando a entrada manual de dados.
- **Extração e Estruturação (Validação de "Contas a Pagar"):** Uso de OCR (Tesseract/Poppler) e *parsers* XML para extrair e normalizar 100% dos dados, preparando-os para o lançamento contábil.
- **Validação Fiscal (A "Primeira Linha de Defesa" do *Compliance*):** Aplicação automática de regras de negócio para validar a consistência de impostos (ICMS, IPI, PIS/COFINS) e totais, bloqueando inconsistências *antes* que elas contaminem o ERP.

- **Armazenamento e Governança (Trilha de Auditoria):** Persistência unificada em um banco de dados PostgreSQL, garantindo um histórico imutável e 100% auditável.
- **Análise Estratégica (O "BI Fiscal"):** Um assistente conversacional (Chat IA) que utiliza RAG (*Retrieval-Augmented Generation*) para permitir que a equipe financeira faça consultas em linguagem natural sobre o acervo, transformando o passivo fiscal em um ativo de dados.

O usuário deixa de ser um digitador e passa a ser um gestor, fazendo o *upload* de lotes de documentos e, em segundos, consultando um agente de IA para obter *insights* e resumos gerenciais.

Imagem 01: Front page da ferramenta.

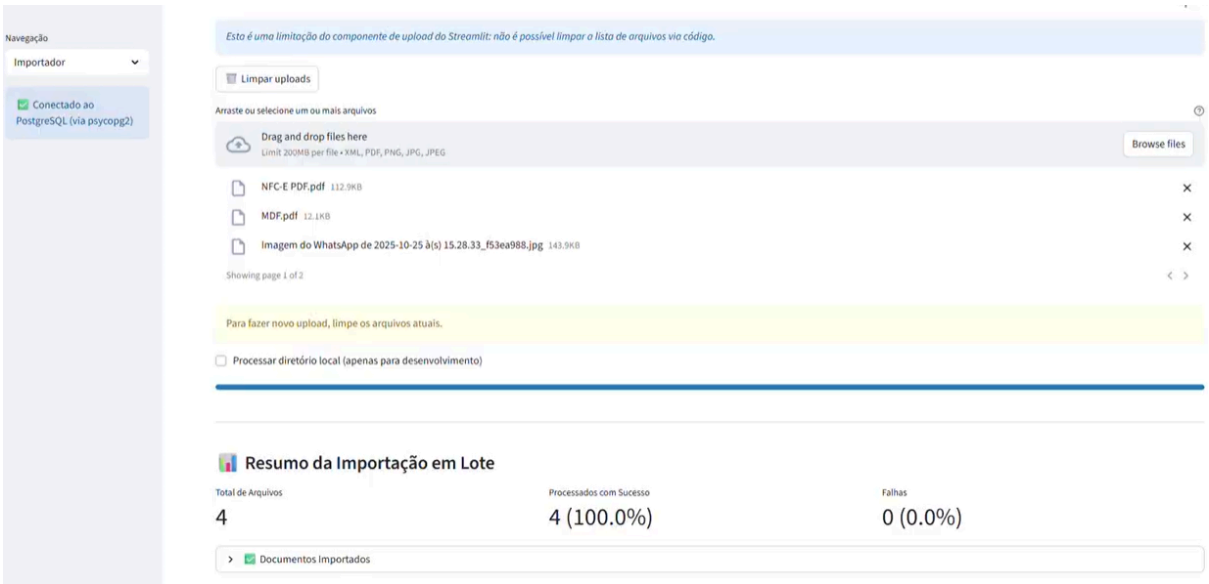


Imagem 02: Importação em lote de documentos.

Imagem 03: Histórico de documentos importados com opção de filtragem.

Imagem 04: Área de Chat-IA Inteligente.

2. Público-Alvo (Quem se beneficia da solução)

A solução foi projetada para gerar valor em diferentes níveis da organização financeira:

- **Analistas Fiscais e Contábeis (Usuários Operacionais):** São os beneficiários diretos da automação. O foco de seu trabalho muda da digitação e verificação manual (baixo valor agregado) para a análise de exceções e planejamento (alto valor agregado).
- **Gestores Financeiros e *Controllers* (Usuários Táticos):** Utilizam a plataforma para garantir a acuracidade do fechamento mensal, acelerar a apuração de impostos, gerar provisões fiscais mais precisas e obter dados consolidados para o planejamento tributário.
- **C-Level (CFOs, VPs de Finanças e Jurídico) (Usuários Estratégicos):** Focados no resultado. Para eles, a plataforma entrega três valores principais: redução de

custos operacionais (OPEX), mitigação de risco de autuações fiscais (passivo contingente) e inteligência de negócio para tomada de decisão.

3. Justificativa do Tema Escolhido (O Valor de Negócio e o ROI)

A gestão fiscal no Brasil é um desafio notório, caracterizado por problemas que representam custos financeiros diretos e indiretos. A ferramenta ataca estes problemas centrais:

1. **Custo Operacional (OPEX) e Ineficiência:** Empresas alocam milhares de *horas-homem* para processar manualmente documentos (NFe, CTe, etc.). Esse custo é agravado pela variedade de formatos (XML, PDF, imagem), exigindo esforço manual dobrado.
2. **Risco de *Compliance* e Custo de Oportunidade:** O processamento manual é um passivo financeiro direto. Erros de digitação ou validação de impostos resultam em pagamentos indevidos (afetando o caixa) ou, pior, em recolhimento a menor, gerando autuações fiscais, multas e juros.
3. **Dados Fiscais como Ativo Morto:** O histórico de documentos fiscais é, na maioria das empresas, um "cemitério de dados". É um repositório valioso que raramente é usado para *insights* estratégicos, representando um custo de armazenamento sem retorno.

O valor agregado do sistema desenvolvido pelo grupo é transformar esse cenário de reativo para proativo, gerando um **ROI claro**:

- **Redução de Custo (OPEX) e Ganho de Produtividade:** A plataforma libera a equipe financeira de tarefas de baixo valor agregado. O tempo de processamento cai de horas para segundos, permitindo que o analista fiscal foque em **análise de créditos tributários**, otimização de rotas logísticas (via CTe) ou negociação com fornecedores (via NFe), em vez de digitação.
- **Mitigação de Risco e Custo de *Compliance***: A validação automática reduz drasticamente o **passivo fiscal contingente** associado a erros humanos. A plataforma cria uma **trilha de auditoria** centralizada e instantânea, facilitando auditorias externas (Big4) e internas, e reduzindo custos com consultorias de remediação.
- **Transformação do Fiscal em *Business Intelligence* (BI):** O Chat IA com RAG desbloqueia o valor dos dados. Um gestor financeiro pode perguntar: "Qual foi nosso gasto total com ICMS-ST de fornecedores de SP no último trimestre?", "Quais fornecedores sistematicamente geram mais créditos de IPI?" ou "Qual a nossa provisão de PIS/COFINS para o próximo fechamento?". A plataforma entrega inteligência de negócio, e não apenas conformidade.

4. Detalhamento do que Foi Desenvolvido (As Ferramentas de Ganho)

A solução é uma aplicação robusta que atua como uma extensão da equipe financeira:

- Importador de Documentos (Garantia de Integridade de Dados):
  O usuário pode processar lotes de arquivos (XML, PDF, JPG, PNG) de uma só vez. O sistema aplica o parser XML ou o pipeline de OCR (Tesseract + Poppler) e, crucialmente, normaliza dados críticos (como formatos numéricos 1.234,56 e datas

DD/MM/YYYY). Isso garante a integridade dos dados antes que eles entrem no ERP, evitando erros no razão contábil.

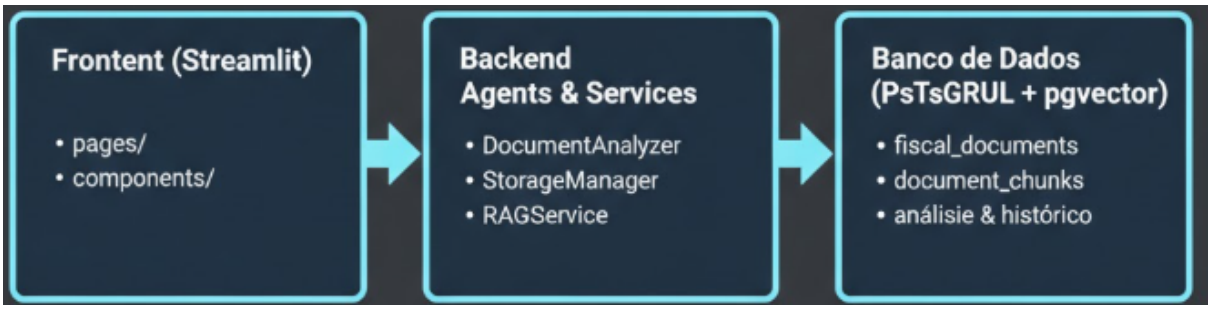
- Guardião de Compliance (Validador Fiscal):
  Após a extração, um agente de validação confere automaticamente a integridade dos dados: checagens de CNPJ/CPF e cálculos de impostos (ICMS, ICMS-ST, IPI, PIS/COFINS). Ele atua como uma primeira linha de defesa; se inconsistências são encontradas, elas são sinalizadas antes do lançamento contábil, evitando a contaminação da base de dados e o retrabalho no fechamento.
- O Controller Assistente (Chat IA com RAG):
  Esta é a funcionalidade central de IA, que democratiza o acesso à informação fiscal.
  - **Exemplo de Uso:** Um gestor solicita um resumo das importações do dia. O agente de IA consulta o banco de dados, analisa os 43 documentos encontrados e retorna um sumário gerencial, agrupado por tipo (NFe, CTe) e por CNPJ do emissor. Isso é um relatório de fechamento diário em segundos.
  - **Exemplo de RAG:** O usuário pede para detalhar as notas de um CNPJ específico. O agente compreende a entidade "CNPJ", executa uma busca filtrada e retorna a lista precisa, uma tarefa que manualmente poderia levar horas de busca em pastas ou no portal da NFe.
  - **Memória Conversacional:** O sistema armazena respostas anteriores, permitindo que o sistema reaproveite análises e aprenda com as interações, garantindo consistência e acelerando respostas futuras.

5. Elementos Adicionais: Arquitetura de Custo Eficiente e Viabilidade de TCO

O diferencial do projeto é o foco no **ROI** da solução. O grupo identificou que o custo de APIs de LLMs (Gemini, OpenAI) seria proibitivo em alto volume. Para mitigar isso, foi implementada uma arquitetura híbrida que garante um **Custo Total de Propriedade (TCO)** baixo e previsível, tornando o *business case* da plataforma viável.

Diagrama de Fluxo da Solução:

(O fluxo demonstra a separação clara entre a interface, os serviços de backend e o banco de dados, que atua como a "memória" do sistema.)



Estratégias de Otimização de Custo (Foco no TCO):

O grupo implementou três estratégias para reduzir o consumo de tokens e o custo de API, garantindo a viabilidade econômica:

Estratégia	Descrição da Implementação	Impacto no Custo (Benefício Financeiro)
1.Pré-Processamento Inteligente	OCR e *parsers* XML extraem e estruturam os dados **antes** de enviá-los ao LLM.	Reduz drasticamente os *tokens* de entrada (custo de *input*). O LLM não precisa "ler" o documento bruto, apenas os dados já limpos.
2. Arquitetura RAG	O banco de dados (PostgreSQL + pgvector) fornece contexto preciso e conciso ao LLM.	Evita *prompts* longos e caros. O LLM recebe apenas a informação relevante (ex: as 3 notas fiscais do fornecedor X), não o banco de dados inteiro.
3. Modelo Híbrido (Otimização de Custo)	Prioriza *embeddings* locais e gratuitos (*Sentence Transformers*) para a busca (RAG).	**Reserva o LLM premium (Gemini) – que é a parte cara – apenas para a tarefa de raciocínio complexo** e geração da resposta final.

Esta arquitetura demonstra uma maturidade financeira e técnica notável. A decisão de usar *embeddings* gratuitos locais (documentada no FREE_EMBEDDINGS_README.md) resolve diretamente o gargalo de custo e quotas de API.

Como detalhado na apresentação, uma análise de custo determinística para 1 milhão de transações/dia com GPT-4o custaria $6.47M anuais, enquanto com Gemini 1.5 Pro custaria $2.56M anuais. A arquitetura híbrida desenvolvida pelo grupo reduz esses custos de forma ainda mais drástica, tornando o produto comercialmente viável para empresas de qualquer porte.

6. Conclusão: De Centro de Custo a Ativo de Inteligência Estratégica

A ferramenta demonstra com sucesso como a aplicação inteligente de Agentes de IA pode redefinir fundamentalmente uma função de *back-office* crítica, como a gestão fiscal. O projeto não se limita a automatizar tarefas manuais — o que, por si só, já representaria um ROI (Retorno sobre o Investimento) claro através da redução de OPEX (Custos Operacionais) e da eliminação de retrabalho.

A verdadeira inovação da plataforma reside em sua dupla capacidade:

1. **Transformação de Valor:** Converte um "passivo" fiscal (um repositório de dados estáticos e de alto risco) em um ativo de inteligência de negócios (BI) dinâmico e consultável. A funcionalidade de RAG (Chat IA) desbloqueia o valor oculto nos

dados, permitindo que gestores tomem decisões estratégicas baseadas em informações que antes eram inacessíveis.

2. **Viabilidade Econômica:** O projeto prova que uma solução de IA avançada pode ser implementada de forma economicamente sustentável. A arquitetura híbrida, que prioriza *embeddings* locais gratuitos para a busca e reserva o LLM premium (Gemini) apenas para o raciocínio complexo, é uma decisão de engenharia madura e focada no negócio. Esta abordagem endereça diretamente o principal gargalo de soluções de IA em larga escala: o Custo Total de Propriedade (TCO).

Em suma, o sistema entrega uma solução completa que atende às três principais demandas do setor financeiro corporativo: reduz custos operacionais, mitiga riscos financeiros e de *compliance* e, o mais importante, gera inteligência acionável para o futuro do negócio.

7. Link para o Repositório do Projeto

O código-fonte completo, documentação técnica, *scripts* de migração e testes da plataforma SkyNET-I2A2 estão disponíveis no seguinte repositório:

Desenvolvido por: SkyNET-I2A2

Github (entrega): https://github.com/efantinatti/SkyNET-I2A2/tree/main/Delivery/20251030

Infográfico: http://efantinatti.github.io/SkyNET-I2A2/Final_Infografico_20251030.html

AWS' Instance App: http://skynet-i2a2.fantinatti.net:8501