1. Descrição do Projeto

O tema escolhido é a extração de dados de documentos fiscais, como notas fiscais eletrônicas, com o objetivo de tornar esse processo mais fácil, rápido e preciso. A ideia é criar uma solução que ajude profissionais e empresas de contabilidade, finanças e tecnologia a coletar, organizar e usar informações importantes de forma automática.

Com essa solução, queremos:

* Recuperar documentos fiscais em fontes conhecidas e confiáveis, tais como a Receita Federal ou prefeituras, seja por upload manual ou download.
* Por meio das tecnologias OCR (reconhecimento óptico de caracteres), inteligência artificial, incluindo modelos de linguagem (LLM), “ler” os documentos e extrair informações como valores, datas, nomes e códigos fiscais.
* Possibilitar que o sistema aprenda com o tempo, adaptando-se a diferentes layouts e formatos de notas fiscais, com o objetivo de impedir que alguma informação não seja coletada. Armazenar essas informações de forma estruturada para que possam ser consultadas ou analisadas a qualquer momento.
* Oferecer uma interface simples onde os usuários possam fazer perguntas e receber respostas claras sobre os dados fiscais.

O foco é economizar tempo, reduzir erros humanos e ajudar empresas, especialmente as que lidam com muitos documentos por dia, como no setor varejista, a cumprir suas obrigações fiscais e realizar auditorias com mais eficiência.

1. Público Alvo

Profissionais e empresas da área de contabilidade, finanças e tecnologia, especialmente aqueles envolvidos com automação de processos fiscais, compliance tributário e extração de dados de documentos fiscais (como notas fiscais eletrônicas).

1. Justificativa para escolha do tema

Empresas que lidam com grandes volumes de notas fiscais precisam extrair dados com rapidez e precisão para cumprir obrigações fiscais e realizar auditorias internas. A adoção de tecnologias como OCR e NLP permite automatizar essa extração, reduzindo erros humanos e custos operacionais. Por exemplo, uma empresa do setor varejista que processa milhares de notas por dia pode economizar centenas de horas de trabalho ao adotar essa solução.

1. História do Usuário

Como profissional da área de contabilidade, finanças ou tecnologia, eu quero uma solução capaz de capturar e extrair automaticamente os dados de documentos fiscais (NF-e), para que eu possa obter essas informações de forma rápida, precisa e organizada, reduzindo erros manuais e otimizando meus processos operacionais.

**Critérios de Aceitação:**

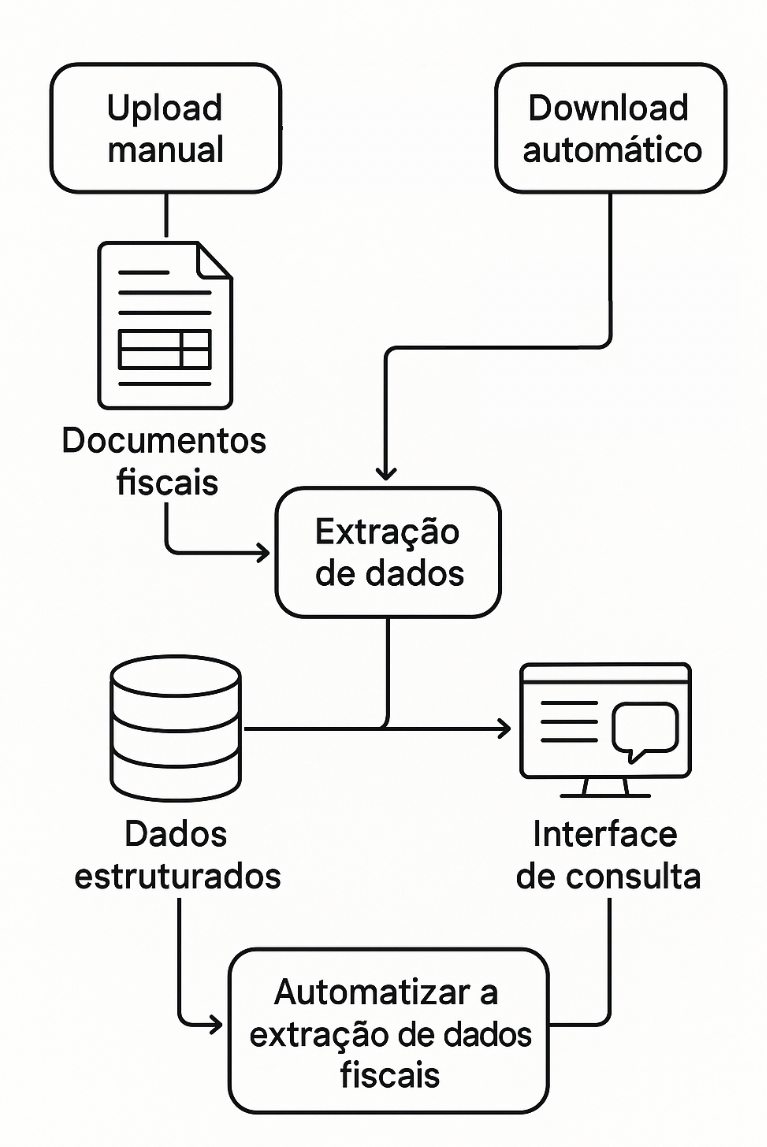
1. O sistema deve permitir o upload manual de documentos fiscais em formato PDF ou imagem.
2. O sistema deve ser capaz de receber uploads ou realizar o download automático de documentos fiscais a partir de fontes oficiais (SEFAZ).
3. A solução deve aplicar tecnologias de OCR e inteligência artificial para extrair dados dos documentos, independentemente de variações no layout.
4. As informações extraídas devem ser armazenadas de forma estruturada e acessível.
5. Deve existir uma interface de interação onde o usuário possa realizar consultas e obter respostas sobre os dados fiscais, como valores, impostos, datas e outros campos relevantes.
6. O sistema deve garantir precisão na extração dos dados, minimizando falhas de reconhecimento ou perdas de informação.

**Objetivo do Negócio:**

Automatizar o processo de extração de dados fiscais, visando:

* Otimização de tempo operacional;
* Redução de erros manuais;
* Maior controle e segurança das informações fiscais;
* Suporte eficiente à tomada de decisão e auditorias.

5. Diagrama



6. Arquitetura da solução



A solução implementa um pipeline automatizado para processamento de documentos fiscais.

1. **Agente 1 (Aquisição de Documentos):** Responsável por obter documentos fiscais (NF-e) em formatos de imagem e PDF, provenientes de upload manual ou download de órgãos governamentais.
2. **Agente 2 (Extração e Aprendizado):** Processa os documentos adquiridos, utilizando OCR para extrair dados e aprender novos layouts com apoio de LLM e dos arquivos armazenados obtidos pelo agente 1, garantindo a extração precisa de informações fiscais relevantes.
3. **Base de Conhecimento:** Utiliza os dados estruturados extraídos pelo Agente 2.
4. **Agente 3 (Resposta e Interação):** Acessa um Large Language Model (LLM) e, utilizando os dados da Base de Conhecimento, responde às perguntas dos usuários sobre as informações fiscais.
5. **Usuário:** Interage com o sistema através de uma interface, recebendo as respostas do Agente 3.

Essa arquitetura permite a automação completa do ciclo de vida dos documentos fiscais, desde a coleta até a disponibilização inteligente das informações.

7. Proposta preliminar do desenvolvimento do projeto

7.1 Arquitetura do Sistema

O sistema será desenvolvido seguindo uma arquitetura multi-agente composta por três agentes especializados:

**Agente 1:** Aquisição de Documentos

Responsabilidade: Obter e pré-processar documentos fiscais

Funcionalidades:

* Interface para upload manual de arquivos (PDF, imagens)
* Integração com APIs de órgãos governamentais (SEFAZ)
* Validação inicial de formato e integridade dos documentos
* Organização e catalogação dos arquivos recebidos

**Agente 2:** Extração e Aprendizado

Responsabilidade: Processar documentos e extrair dados relevantes

Funcionalidades:

* OCR avançado para digitalização de documentos
* NLP para identificação e extração de campos específicos
* IA para adaptação a novos layouts
* Validação cruzada de dados extraídos

**Agente 3:** Resposta e Interação

Responsabilidade: Interface inteligente com usuários

Funcionalidades:

* Integração com LLMs para consultas em linguagem natural

7.2 Componentes de Suporte

**Fontes de Documentos**

Upload manual via interface web

Integração com sistemas governamentais

**Armazenamento e Acesso a LLMs**

Banco de dados SQLite para documentos processados

Cache inteligente para otimização de performance

Integração com APIs de LLMs (Gemini)

**Base de Conhecimento**

Templates de documentos fiscais brasileiros

7.3 Tecnologias Principais

**Backend e Processamento**

Pythoncomo linguagem principal

Urllib3 para consumir APIs

Langchain como biblioteca de interação com as LLMs

SQLitepara documentos e metadados

**Inteligência Artificial**

Tesseract OCR+ OpenCV para reconhecimento de texto

GeminiAIpara capacidades avançadas de LLM

**Frontend e Interface**

React.js

Tailwind CSS para estilização responsiva

Lucide React para ícones (Send, Bot, User, Loader2, etc.)

Fetch API para chamadas HTTP ao LangChain

**Infraestrutura**

GitHub para versionamento e divulgação do código

7.4 Metodologia de Desenvolvimento

**Fase 1: Levantamento de Requisitos – Primeira Apresentação**

* Análise detalhada de tipos de documentos fiscais brasileiros
* Definição de campos obrigatórios para extração
* Estudo de regulamentações fiscais aplicáveis
* Coleta de amostras de documentos para treinamento (Agente 2)

**Fase 2: Desenvolvimento OCR + NLP – Agente 2 – Primeira Apresentação**

* Implementação do motor de OCR otimizado
* Desenvolvimento de algoritmos de NLP para extração de campos
* Criação de templates para diferentes tipos de documentos e aprendizado

**Fase 3: Testes Iniciais – Primeira Apresentação**

* Validação com conjunto de documentos de teste
* Ajustes de precisão e correção de bugs
* Documentação técnica inicial

**Fase 4: Atividades Finais – Primeira Apresentação**

* Demonstração do MVP funcional
* Apresentação de métricas de precisão alcançadas
* Planejamento de melhorias

**Fase 5: Interface – Agente 1 + Frontend – Segunda Apresentação**

* Desenvolvimento de APIs para download de arquivos
* Criação da interface web de usuário
* Implementação de funcionalidades avançadas
* Testes de download de dados com sites reais

**Fase 6: Atividades Finais – Segunda Apresentação**

* Demonstração do sistema completo
* Apresentação de casos de uso reais
* Análise de ROI e benefícios mensuráveis
* Roadmap para desenvolvimento futuro

**Fase 7: Validação e Ajustes – Terceira Apresentação**

* Testes com usuários finais
* Correções e otimizações finais
* Preparação da documentação completa
* Validação de compliance

**Fase 8: Apresentação final Intermediária - Terceira Apresentação**

* Apresentação final do projeto
* Entrega de código-fonte e documentação
* Demonstração de impacto e resultados
* Plano de continuidade e manutenção

**Fase 9: Apresentação Final (Parte 1)**

* Entrega de código-fonte e documentação

**Fase 10: Apresentação Final (Parte 2)**

* Entrega de código-fonte e documentação

7.5 Métricas de Sucesso

**Técnicas**

Precisão de extração: > 95% para campos obrigatórios

Tempo de processamento: < 30 segundos por documento

Taxa de falsos positivos: < 3%

Disponibilidade do sistema: > 99.5%

**Negócio**

Redução de tempo de processamento: > 80%

Economia de custos operacionais: > R$ 50.000/ano

Satisfação do usuário: > 4.5/5.0

ROI do projeto: > 300% no primeiro ano

7.6 Desafios Identificados e Soluções

**Desafio 1:** Variedade de Layouts

Problema: Documentos fiscais possuem layouts diversos

Solução: Machine Learning adaptativo e templates dinâmicos

**Desafio 2:** Qualidade de Imagens

Problema: Documentos escaneados com baixa qualidade

Solução: Pré-processamento avançado de imagens e OCR robusto

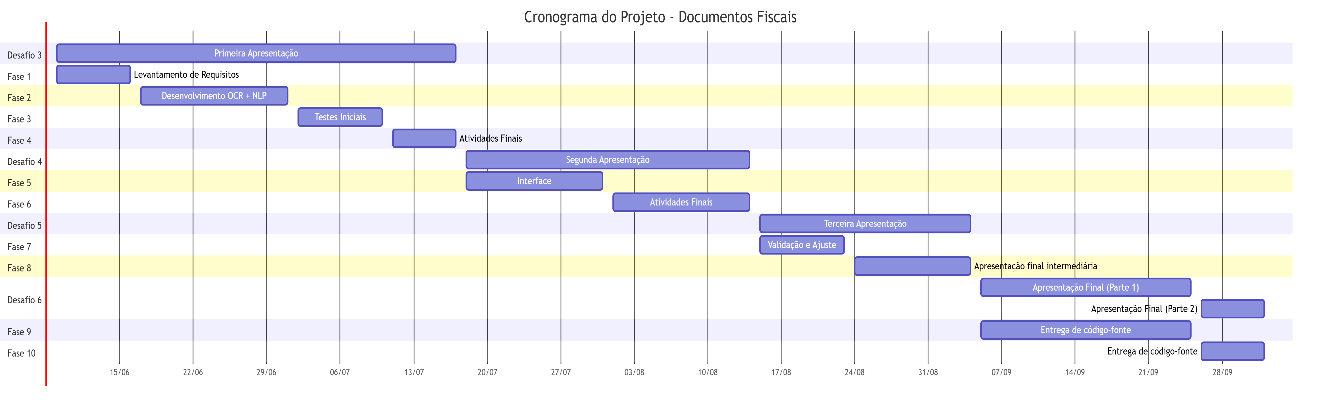
**Desafio 3:** Compliance Legal

Problema: Adequação às normas fiscais brasileiras

Solução: Consultoria especializada e validação jurídica contínua

8. Cronograma

|  |  |  |  |
| --- | --- | --- | --- |
| **Fase** | **Atividades Principais** | **Período Estimado** | **Duração** |
|  | **Desafio 3 – Primeira Apresentação** | **09/06 à 17/07** | |
| Fase 1 | **Levantamento de Requisitos**  - Análise detalhada de tipos de documentos fiscais brasileiros - Definição de campos obrigatórios para extração - Estudo de regulamentações fiscais aplicáveis - Coleta de amostras de documentos para treinamento | 09/06 à 16/06 | 8 dias |
| Fase 2 | **Desenvolvimento OCR + NLP - Agente 2**  - Implementação do motor de OCR otimizado - Desenvolvimento de algoritmos de NLP para extração de campos - Criação de templates para diferentes tipos de documentos e aprendizado. | 17/06 à 01/07 | 15 dias |
| Fase 3 | **Testes Iniciais**  - Elaboração de plano detestes  - Validação com conjunto de documentos de teste - Ajustes de precisão e correção de bugs - Documentação técnica inicial | 02/07 à 10/07 | 9 dias |
| Fase 4 | **Atividades Finais** - Planejamento de melhorias  - Apresentação de métricas de precisão alcançadas - Demonstração do MVP funcional | 11/07 à 17/07 | 7 dias |
|  | **Desafio 4 – Segunda Apresentação** | **18/07 à 14/08** | |
| Fase 5 | **Interface - Agente 1 + Frontend** - Desenvolvimento de APIs para download de arquivos - Criação de interface web - Implementação de funcionalidades avançadas - Testes de download de dados com sites reais | 18/07 à 31/07 | 14 dias |
| Fase 6 | **Atividades Finais**  - Demonstração do sistema completo - Casos reais e ROI - Roadmap futuro | 01/08 à 14/08 | 14 dias |
|  | **Desafio 5 – Terceira Apresentação** | **15/08 à 04/09** | |
| Fase 7 | **Validação e Ajuste** | 15/08 à 23/08 | 9 dias |
| - Testes com usuários finais - Correções e otimizações finais - Preparação da documentação completa - Validação de compliance |
| Fase 8 | **Apresentação final intermediária** - Entrega de código-fonte e documentação - Demonstração de impacto e resultados - Plano de continuidade e manutenção | 24/08 à 04/09 | 12 dias |
|  | **Desafio 6 – Apresentação Final (Parte 1)** | **05/09 à 25/09** | |
| Fase 9 | - Entrega de código-fonte e documentação | 05/09 à 25/09 | 21 dias |
|  | **Desafio 6 – Apresentação Final (Parte 2)** | **26/09 à 02/10** | |
| Fase 10 | - Entrega de código-fonte e documentação | 26/09 à 02/10 | 7 dias |

9. Linha do Tempo

10. Requisitos de Dados

**RD001: Análise Detalhada de Tipos de Documentos Fiscais Brasileiros** **:**

O sistema deve ser capaz de processar, no mínimo, os seguintes documentos:

* **NF-e (Nota Fiscal Eletrônica):** Documento para circulação de mercadorias. Padrão nacional, mas com variações de layout.
* **NFS-e (Nota Fiscal de Serviços Eletrônica):**

Documento para prestação de serviços. **Grande desafio:** não há um padrão nacional; cada prefeitura tem seu próprio layout e sistema.

* **CT-e (Conhecimento de Transporte Eletrônico):**

Documento para serviços de transporte de cargas.

**RD002: Definição de Campos Obrigatórios para Extração.**

A seguir, uma lista de campos essenciais a serem extraídos. O modelo de IA deve ser treinado para encontrá-los.

|  |  |  |  |
| --- | --- | --- | --- |
| **Categoria** | **Campo** | **Exemplo** | **Tipo de Documento** |
| Geral | Número da Nota | 12345 | Todos |
|  | Série | 1 | Todos |
|  | Data de Emissão/Competência | 07/06/2025 | Todos |
|  | Chave de Acesso | 412... (44 dígitos) | NF-e, CT-e |
|  | Código de Verificação | ABC123XYZ | NFS-e |
| Emitente | CNPJ/CPF | 00.123.456/0001-78 | Todos |
|  | Razão Social | Exemplo LTDA | Todos |
|  | Endereço, Município, UF | Rua Exemplo, 123, São Paulo, SP | Todos |
| Destinatário | CNPJ/CPF | 98.765.432/0001-10 | Todos |
|  | Razão Social | Cliente Final S.A. | Todos |
| Valores | Valor Total da Nota | R$1.500,00 | Todos |
|  | Valor Total dos Produtos/Serviços | R$1.450,00 | Todos |
|  | Base de Cálculo do ICMS | R$1.450,00 | NF-e, CT-e |
|  | Valor do ICMS | R$261,00 | NF-e, CT-e |
|  | Valor do PIS/COFINS | R$50,00 | Todos |
|  | Valor do ISS | R$72,50 | NFS-e |
| Itens | Descrição do Produto/Serviço | Consultoria, Parafuso Sextavado | NF-e, NFS-e |
|  | Quantidade | 10 | NF-e, NFS-e |
|  | Valor Unitário | R$145,00 | NF-e, NFS-e |
|  | Código Fiscal (NCM/CNAE) | 8599-6/04, 7490-1/04 | NF-e, |

**RD003: Estudo de Regulamentações Fiscais Aplicáveis**

* **LGPD (Lei Geral de Proteção de Dados):** O sistema lidará com dados sensíveis (CNPJ, valores). É **mandatório** que o sistema seja projetado em conformidade com a LGPD, garantindo a segurança, o consentimento (quando aplicável) e a privacidade dos dados.
* **Legislação Tributária:** O sistema não precisa ser um emissor ou validador fiscal, mas o conhecimento das regras (ex: o que é NCM, como o ICMS é calculado) é crucial para a correta extração e categorização dos dados. A equipe deve ter acesso a um consultor fiscal ou material de referência.

11. Requisitos Funcionais

**Módulo 1: Captura e Processamento de Documentos**

**RF001: Upload Manual de Documentos:**

* O sistema deve permitir que o usuário faça o upload de um ou mais documentos fiscais simultaneamente.
* Formatos suportados: PDF (texto e imagem), JPG, PNG.
* Deve haver uma fila de processamento visível ao usuário (Ex: "Aguardando", "Processando", "Concluído", "Erro").

**RF002: Download Automático de Fontes Oficiais.**

* O sistema deve ter a capacidade de se conectar a portais de Secretarias da Fazenda (SEFAZ) e prefeituras para baixar documentos fiscais.
* Isso exigirá que o usuário configure suas credenciais de acesso (ex: certificado digital A1/A3, login e senha).
* O sistema deve realizar buscas periódicas e automáticas por novos documentos.

**RF003: Processamento de OCR:**

* Para documentos baseados em imagem (PDF de imagem, JPG, PNG), o sistema deve aplicar uma tecnologia de Reconhecimento Óptico de Caracteres (OCR) para converter a imagem em texto bruto.

**RF004: Extração Inteligente de Dados (IA/LLM):**

O sistema utilizará modelos de IA e LLM para:

* Identificar o tipo de documento (NF-e, NFS-e, CT-e, etc.).
* Analisar o texto extraído e identificar os campos-chave (CNPJs, datas, valores, etc.), mesmo que estejam em posições diferentes em layouts variados.
* Calcular um "score de confiança" para cada campo extraído. Campos com baixa confiança devem ser sinalizados para revisão humana.

**RF005: Interface de Validação e Correção:**

* O usuário deve ter uma tela onde possa visualizar o documento original lado a lado com os dados extraídos.
* Nesta tela, ele poderá corrigir ou preencher campos que a IA não conseguiu extrair ou extraiu incorretamente.
* As correções feitas pelo usuário devem ser usadas como feedback para o sistema (ver RF007).

**Módulo 2: Armazenamento e Aprendizado**

**RF006: Armazenamento Estruturado:**

* Os dados extraídos devem ser salvos em um banco de dados de forma estruturada.
* Cada registro de dados deve manter um vínculo com o arquivo original do documento.

**RF007: Mecanismo de Aprendizado Contínuo (Machine Learning):**

* O sistema deve usar as correções manuais (RF005) para retroalimentar e refinar os modelos de extração de dados, melhorando sua precisão ao longo do tempo para novos layouts e formatos.

**Módulo 3: Consulta e Análise**

**RF008: Interface de Consulta em Linguagem Natural:**

* O sistema deve oferecer um chat onde o usuário possa fazer perguntas como:

"Qual o total gasto com o fornecedor X no último mês?"

"Liste todas as notas fiscais de serviço de São Paulo em maio."

"Quanto pagamos de ICMS em 2024?"

**RF009: Geração de Respostas Claras:**

* O LLM deve interpretar a pergunta, buscar os dados no banco de dados estruturado e apresentar a resposta de forma clara e objetiva, podendo ser em texto, tabelas ou valores numéricos.

**RF010: Dashboard e Relatórios:**

* O sistema deve apresentar um painel visual (dashboard) com os principais indicadores (total faturado, principais fornecedores, impostos por período, etc.).

**RF011: Exportação de Dados:**

* O usuário deve ser capaz de exportar os resultados de suas buscas ou relatórios para formatos comuns (CSV, Excel).

12. Requisitos Não Funcionais:

**RNF001: Segurança**

* **Controle de Acesso:** Autenticação de usuários baseada em login e senha, com políticas de senha fortes.
* **Criptografia:** Todos os dados sensíveis devem ser criptografados, tanto em trânsito (HTTPS/SSL) quanto em repouso (no banco de dados e no armazenamento de arquivos).
* **Gestão de Credenciais:** As credenciais para acesso aos portais do governo (certificados, senhas) devem ser armazenadas de forma extremamente segura.

**RNF002: Desempenho:**

* O processamento de um único documento (upload, OCR, extração) deve levar no máximo 60 segundos.
* As consultas em linguagem natural devem retornar respostas em menos de 5 segundos para a maioria das perguntas.

**RNF003: Usabilidade:**

* A interface deve ser limpa, intuitiva e exigir o mínimo de treinamento para o usuário final.
* As mensagens de erro devem ser claras e orientar o usuário sobre como proceder.

**RNF 004: Escalabilidade:**

* A arquitetura deve ser projetada para suportar um aumento no número de usuários e no volume de documentos sem degradação significativa do desempenho (ex: uso de micro serviços, filas e recursos de nuvem auto escaláveis).

**RNF 005: Disponibilidade:**

* O sistema deve ter uma disponibilidade de pelo menos 99.5% (uptime).

11. Integrantes do Grupo

* Ana Paola Zarate Zequita
* Antonio João Nascimento Dantas
* Aurélio Marques Morbeck
* Daniel Correa Rodrigues
* Eduarda Silveira Cardoso
* Lorrane da F C Santos Santana
* Luciane Schumacher
* Matheus de Oliveira Vieira
* Valdinei de Souza Camargo
* Walison Carlos da Silva