**Architecture Decision Records (ADRs)**

**ADR-001: Escolha da Arquitetura Híbrida (SaaS + Desktop Agent)**

**Status**: Aprovado  
**Data**: 2025-01-20  
**Decidores**: Equipe de Arquitetura, Product Owner

**Contexto**

Precisamos decidir entre desenvolver uma aplicação desktop (como sugerido inicialmente) ou uma plataforma SaaS, considerando as necessidades específicas do mercado jurídico brasileiro.

**Decisão**

Adotaremos uma **arquitetura híbrida** com foco principal em SaaS e um componente desktop opcional para casos específicos.

**Justificativa**

**Prós da Solução Híbrida:**

* **Acessibilidade**: SaaS permite acesso de qualquer lugar/dispositivo
* **Colaboração**: Facilita trabalho em equipe e compartilhamento
* **Manutenção**: Atualizações centralizadas e automáticas
* **Escalabilidade**: Crescimento natural com demanda
* **Privacidade**: Desktop agent para dados ultra-sensíveis
* **Performance**: IA local para processamento intensivo

**Contras Mitigados:**

* **Dependência de Internet**: Modo offline básico no SaaS + desktop agent
* **Controle de Dados**: Criptografia end-to-end + opção de IA local
* **Performance**: Caching inteligente + processamento local quando necessário

**Implicações**

* Desenvolvimento em duas frentes: web-first, desktop complementar
* Arquitetura de APIs robusta para sincronização
* Estratégia de dados híbrida (cloud + local)

**ADR-002: Escolha do Stack Tecnológico Principal**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

**Frontend SaaS**: React 18 + TypeScript + Next.js 14 + TailwindCSS  
**Backend**: Node.js + NestJS + PostgreSQL + Redis  
**Desktop**: Electron + React  
**IA**: LangChain + OpenAI/Anthropic APIs + Ollama (local)  
**Infraestrutura**: AWS/Azure + Kubernetes

**Justificativa**

* **React/TypeScript**: Ecossistema maduro, produtividade, type safety
* **Next.js**: SSR, performance, developer experience
* **NestJS**: Arquitetura modular, TypeScript native, decorators
* **PostgreSQL**: ACID compliance, JSON support, extensibilidade
* **Electron**: Reutilização de código web, acesso a recursos locais

**ADR-003: Estratégia Anti-Alucinação da IA**

**Status**: Aprovado  
**Data**: 2025-01-20

**Contexto**

Baseado no caso real citado (juiz Matheus de Lima Sampaio criticando uso inadequado de IA), precisamos garantir veracidade total das informações geradas.

**Decisão**

Implementaremos um **Sistema Anti-Alucinação Multicamadas**:

1. **RAG Obrigatório**: Toda resposta baseada em documentos verificáveis
2. **Validação de Citações**: Verificação automática de todas as referências
3. **Confidence Scoring**: Score de confiança para cada resposta
4. **Human-in-the-Loop**: Aprovação humana obrigatória para documentos críticos
5. **Transparência Total**: Trilha de auditoria para cada sugestão

**Implicações**

* Performance pode ser impactada pela validação rigorosa
* Necessidade de base de conhecimento jurídica extensa e atualizada
* Interface deve comunicar claramente limitações e incertezas

**ADR-004: Modelo de Dados para Compliance LGPD**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Implementaremos **Privacy by Design** com:

* Criptografia AES-256 para dados sensíveis
* Pseudonimização automática de dados pessoais
* Crypto-shredding para "direito ao esquecimento"
* Logs imutáveis de auditoria
* Gestão granular de consentimentos

**Justificativa**

LGPD é requisito legal incontornável. Implementar desde o início é mais eficiente que retrofit posterior.

**ADR-005: Estratégia de Integração com Tribunais**

**Status**: Em Revisão  
**Data**: 2025-01-20

**Contexto**

Necessidade de integração com sistemas de tribunais (PJe, e-SAJ, Projudi, e-proc) para automação de consultas e protocolo.

**Opções Consideradas**

1. **APIs Oficiais**: Quando disponíveis
2. **Web Scraping Ético**: Com respeito aos termos de uso
3. **RPA (Robotic Process Automation)**: Automação de interface
4. **Parcerias**: Com provedores de dados jurídicos

**Decisão Provisória**

Abordagem híbrida:

* Priorizar APIs oficiais quando disponíveis
* Web scraping ético como fallback
* RPA assistido (não autônomo) para protocolo
* Parcerias estratégicas para dados de qualidade

**Riscos**

* Mudanças nos sistemas dos tribunais
* Questões legais sobre scraping
* Rate limiting e bloqueios

**ADR-006: Arquitetura de Microserviços**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Adotaremos arquitetura de microserviços com os seguintes serviços principais:

1. **Auth Service**: Autenticação e autorização
2. **Document Service**: Gestão de documentos e templates
3. **AI Engine Service**: Motor de IA e RAG
4. **Legal Database Service**: Base de conhecimento jurídico
5. **Deadline Service**: Gestão de prazos
6. **Case Management Service**: Gestão de casos
7. **Court Integration Service**: Integrações com tribunais
8. **Notification Service**: Sistema de notificações
9. **Analytics Service**: Business intelligence

**Justificativa**

* Escalabilidade independente por serviço
* Deploy independente
* Tecnologias específicas por domínio
* Resiliência (falha isolada)
* Facilita testes e manutenção

**Trade-offs**

* Complexidade de deployment
* Latência de comunicação inter-serviços
* Necessidade de service mesh
* Debugging mais complexo

**ADR-007: Estratégia de Caching**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Implementaremos caching em múltiplas camadas:

1. **Browser Cache**: Assets estáticos
2. **CDN**: Conteúdo estático global
3. **Application Cache**: Redis para dados frequentes
4. **Database Cache**: PostgreSQL query cache
5. **RAG Cache**: Embeddings e resultados de busca

**Configuração**

Browser: 1 hora (assets versionados)

CDN: 24 horas (assets estáticos)

Redis: TTL variável por tipo de dado

RAG Results: 6 horas (jurisprudência muda lentamente)

User Sessions: 24 horas

**ADR-008: Estratégia de Testes**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Pirâmide de testes com foco em qualidade da IA:

1. **Unit Tests**: 70% cobertura mínima
2. **Integration Tests**: APIs e banco de dados
3. **AI Quality Tests**: Golden datasets para IA
4. **E2E Tests**: Jornadas críticas do usuário
5. **Load Tests**: Performance sob carga
6. **Security Tests**: OWASP compliance

**Ferramentas**

* **Jest**: Unit e integration tests
* **Playwright**: E2E tests
* **K6**: Load testing
* **OWASP ZAP**: Security testing
* **Custom**: Validação de qualidade da IA

**ADR-009: Monitoramento e Observabilidade**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Stack de observabilidade completo:

* **Metrics**: Prometheus + Grafana
* **Logs**: ELK Stack (Elasticsearch, Logstash, Kibana)
* **Tracing**: Jaeger para request tracing
* **APM**: New Relic ou Datadog
* **Uptime**: Pingdom ou StatusPage
* **Alerts**: PagerDuty integration

**KPIs Críticos**

* Uptime > 99.9%
* Response time < 2s (95th percentile)
* AI generation time < 30s
* Citation validation accuracy > 99.5%
* User satisfaction score > 4.5/5

**ADR-010: Estratégia de Segurança**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Implementaremos segurança em múltiplas camadas:

**Autenticação e Autorização**

* OAuth 2.0 + OpenID Connect
* MFA obrigatório para produção
* JWT com refresh tokens
* RBAC granular por recurso

**Criptografia**

* TLS 1.3 em trânsito
* AES-256 em repouso
* Chaves gerenciadas via HashiCorp Vault
* Crypto-shredding para LGPD

**Aplicação**

* Input validation e sanitization
* SQL injection prevention (ORM + prepared statements)
* XSS protection (CSP headers)
* CSRF tokens
* Rate limiting por IP e usuário

**Infraestrutura**

* WAF (Web Application Firewall)
* DDoS protection
* Network segmentation
* Regular security scans
* Penetration testing quarterly

**ADR-011: Estratégia de Backup e Disaster Recovery**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Implementaremos estratégia 3-2-1:

* **3** cópias dos dados
* **2** tipos diferentes de mídia
* **1** cópia offsite

**Configuração**

* **Backup Incremental**: Diário às 02:00 UTC
* **Backup Full**: Semanal aos domingos
* **Replicação Cross-Region**: Tempo real para dados críticos
* **Point-in-Time Recovery**: Até 30 dias atrás
* **Testes de Restore**: Mensais automatizados

**SLA de Recovery**

* **RTO** (Recovery Time Objective): 4 horas
* **RPO** (Recovery Point Objective): 1 hora
* **Disponibilidade**: 99.9% anual

**ADR-012: Licenciamento de Modelos de IA**

**Status**: Em Análise  
**Data**: 2025-01-20

**Contexto**

Necessidade de definir quais modelos de IA utilizar, considerando custos, performance e questões legais.

**Opções**

1. **Modelos Proprietários** (GPT-4, Claude)
   * Prós: Alta qualidade, suporte comercial
   * Contras: Custo alto, dependência externa
2. **Modelos Open Source** (Llama, Mistral)
   * Prós: Controle total, sem custos de API
   * Contras: Necessita infraestrutura própria
3. **Modelos Fine-tuned**
   * Prós: Especialização jurídica
   * Contras: Custo e complexidade de treinamento

**Decisão Provisória**

Abordagem híbrida:

* **Produção**: GPT-4o para geração crítica
* **Local**: Llama 3.1 8B para casos sensíveis
* **Fine-tuning**: Modelo específico para classificação jurídica

**ADR-013: Estratégia de Base de Conhecimento Jurídico**

**Status**: Aprovado  
**Data**: 2025-01-20

**Decisão**

Construiremos base de conhecimento própria com:

**Fontes Primárias**

* **Legislação**: Planalto, Senado, Câmara, Assembleis
* **Jurisprudência**: Portais dos tribunais via APIs/scraping
* **Súmulas**: STF, STJ, TST compiladas
* **Temas Repetitivos**: STJ e STF

**Pipeline de Atualização**

1. **Crawling Automático**: Daily para novidades
2. **Processamento**: Extração + estruturação
3. **Validação**: Quality check automatizado
4. **Indexação**: Embeddings + metadados
5. **Deploy**: Hot swap sem downtime

**Qualidade dos Dados**

* Validação cruzada entre fontes
* Versionamento com controle de mudanças
* Métricas de qualidade por fonte
* Human review para conteúdo crítico

**ADR-014: Modelo de Pricing e Quotas**

**Status**: Em Discussão  
**Data**: 2025-01-20

**Contexto**

Definir modelo de negócio e sistema de quotas para uso de IA.

**Opções Consideradas**

1. **Freemium**: Gratuito limitado + planos pagos
2. **Assinatura Fixa**: Mensalidade com tudo incluso
3. **Pay-per-Use**: Cobrança por documento gerado
4. **Híbrido**: Base fixa + extras por uso

**Proposta Atual**

**Modelo Híbrido** com 4 tiers:

**Tier 1 - Solo (R$ 97/mês)**

* 50 documentos/mês com IA
* Até 3 casos ativos
* Prazos ilimitados
* Pesquisa básica

**Tier 2 - Escritório Pequeno (R$ 297/mês)**

* 200 documentos/mês com IA
* Até 3 usuários
* Casos ilimitados
* Pesquisa avançada
* Integrações básicas

**Tier 3 - Escritório Médio (R$ 597/mês)**

* 500 documentos/mês com IA
* Até 10 usuários
* Analytics avançado
* API access
* Suporte prioritário

**Tier 4 - Enterprise (Custom)**

* Documentos ilimitados
* Usuários ilimitados
* IA local opcional
* Customizações
* SLA dedicado

**Quotas e Rate Limiting**

* **Documentos IA**: Por tier mensal
* **API Calls**: 1000/hour por usuário
* **Storage**: 100GB base + R$ 10/100GB extra
* **Processos Monitorados**: 100 base + R$ 1/processo extra