Canvas de Planejamento de Escalabilidade

Sistema de Predição Automatizada de Maturidade Óssea

---

1. Objetivo da Escalabilidade

Propósito Principal: Capacitar o sistema de predição de idade óssea para atender à demanda crescente de múltiplos laboratórios e instituições médicas, mantendo alta performance, precisão diagnóstica e disponibilidade 24/7. O objetivo é evoluir de um sistema piloto (Laboratório José Rocha de Sá) para uma plataforma SaaS nacional, suportando milhares de exames simultâneos sem degradação da qualidade do serviço.

Metas de Escalabilidade:

- Expandir de 1 laboratório para 10-20 instituições médicas
- Suportar processamento de 1.000+ exames/dia (vs. 50-100 atuais)
- Manter tempo de resposta < 3 segundos mesmo com alta concorrência
- Garantir 99,9% de uptime para operação crítica em saúde

---

2. Volume Esperado de Interações

Cenário Atual (Piloto - José Rocha de Sá):

- **Exames/dia:** 50-100 radiografias
- **Pico horário:** 10-15 exames/hora (manhã)
- **Usuários simultâneos:** 2-5 radiologistas
- **Requisições/API:** ~200-300/dia
- **Armazenamento:** ~500MB-1GB/dia

Cenário Escalado (6 meses):

- **Exames/dia:** 1.000-2.000 radiografias
- **Pico horário:** 150-200 exames/hora
- **Usuários simultâneos:** 20-50 radiologistas
- **Requisições/API:** ~5.000-8.000/dia
- **Armazenamento:** ~10-20GB/dia

Cenário Futuro (12-24 meses):

- **Exames/dia:** 5.000+ radiografias
- **Pico horário:** 500+ exames/hora
- **Usuários simultâneos:** 100+ radiologistas
- **Requisições/API:** ~25.000+/dia
- **Armazenamento:** ~50GB+/dia

---

3. Requisitos de Infraestrutura

Computação:

- **CPU:** Mínimo 16 cores para processamento de ML
- **GPU:** NVIDIA V100/A100 para inferência otimizada
- **RAM:** 64GB+ para múltiplos modelos em memória
- **Container Orchestration:** Kubernetes para auto-scaling

Armazenamento:

- **Volume Persistente:** 10TB+ SSD para modelos e cache
- **Backup:** S3/Cloud Storage para imagens processadas
- **CDN:** Para distribuição global de assets estáticos

Rede:

- **Bandwidth:** 10Gbps+ para upload simultâneo de imagens
- **Load Balancer:** Nginx/HAProxy para distribuição de carga
- **API Gateway:** Rate limiting e autenticação centralizados

Banco de Dados:

- **Primário:** PostgreSQL para metadados e usuários
- **Cache:** Redis para resultados frequentes
- **Analytics:** InfluxDB para métricas de performance

---

4. Estratégias de Escalabilidade

Horizontal Scaling:

- **Microserviços:** Separação de processamento, inferência e API
- **Auto-scaling:** Kubernetes HPA baseado em CPU/memória
- **Queue System:** Redis/RabbitMQ para processamento assíncrono

- **Stateless Design:** APIs sem estado para facilitar replicação

Otimização de Modelo:

- **Model Quantization:** Redução de precisão (FP16) para velocidade
- **Batch Inference:** Processamento de múltiplas imagens simultaneamente
- **Model Caching:** Cache inteligente de predições similares
- **GPU Sharing:** Múltiplos workers compartilhando recursos GPU

Arquitetura de Dados:

- **Database Sharding:** Particionamento por região/laboratório
- **Read Replicas:** Múltiplas cópias para consultas
- **Data Lifecycle:** Arquivamento automático de dados antigos

Multi-Region Deployment:

- **Geo-Distribution:** Nodes regionais (SP, RJ, PE, BA)
- **Edge Computing:** Inferência próxima aos usuários
- **Data Residency:** Conformidade com regulamentações locais

---

5. Custo Estimado

Infraestrutura Cloud (Mensal):

Componente	Atual	6 meses	12 meses
Compute (GPU)	R$ 2.000	R$ 8.000	R$ 20.000
Storage	R$ 500	R$ 2.000	R$ 5.000
Network/CDN	R$ 300	R$ 1.200	R$ 3.000
Database	R$ 800	R$ 2.500	R$ 6.000
Monitoring	R$ 200	R$ 500	R$ 1.000
Total	**R$ 3.800**	**R$ 14.200**	**R$ 35.000**

Desenvolvimento e Operação:

- **DevOps Engineer:** R$ 12.000/mês
- **ML Engineer:** R$ 15.000/mês
- **Support/SRE:** R$ 8.000/mês (após 6 meses)

ROI Projetado:

- **Revenue 6 meses:** R$ 50.000/mês (10 clientes × R$ 5.000)
- **Revenue 12 meses:** R$ 150.000/mês (30 clientes × R$ 5.000)
- **Break-even:** 4-5 meses após início da escalabilidade

---

6. Riscos e Mitigação

Riscos Técnicos:

Risco	Probabilidade	Impacto	Mitigação
Latência aumentada	Alta	Alto	Cache inteligente, otimização de modelo
Falha de GPU	Média	Alto	Redundância, fallback para CPU
Overload de API	Alta	Médio	Rate limiting, circuit breakers
Corrupção de dados	Baixa	Alto	Backup automático, versionamento

Riscos de Negócio:

Risco	Probabilidade	Impacto	Mitigação
Regulamentação médica	Média	Alto	Certificação ANVISA, auditoria contínua
Competição	Alta	Médio	Diferenciação por precisão e suporte
Churn de clientes	Média	Alto	SLA robusto, suporte 24/7
Custos acima do previsto	Alta	Médio	Monitoramento rigoroso, alertas de budget

Riscos Operacionais:

Risco	Probabilidade	Impacto	Mitigação
Falta de expertise	Média	Alto	Treinamento da equipe, consultoria externa
Dependência de fornecedor	Baixa	Alto	Multi-cloud strategy, vendor diversification

Security breach	Baixa	Alto	Encryption end-to-end, auditorias regulares

---

7. Monitoramento de Escalabilidade

Métricas de Performance:

- **Latência P95:** < 3 segundos para 95% das requisições
- **Throughput:** Requisições/segundo por node
- **Error Rate:** < 0.1% de falhas
- **Resource Utilization:** CPU, GPU, Memory por serviço

Métricas de Negócio:

- **MAE do Modelo:** Manter ≤ 10 meses independente da escala
- **Uptime:** 99.9% de disponibilidade
- **Customer Satisfaction:** NPS ≥ 8.0
- **Time to Market:** Onboarding de novos clientes < 48h

Ferramentas de Monitoramento:

- **APM:** Datadog/New Relic para performance end-to-end
- **Infrastructure:** Prometheus + Grafana para métricas de sistema
- **Logs:** ELK Stack para agregação e análise
- **Alerting:** PagerDuty para incidentes críticos
- **Business Intelligence:** Tableau para métricas de negócio

Dashboards Críticos:

1. **Operations Dashboard:** Uptime, latência, error rates
2. **ML Model Dashboard:** Accuracy, drift detection, prediction distribution
3. **Business Dashboard:** Revenue, churn, usage patterns
4. **Cost Dashboard:** Spend por cliente, otimização de recursos

---

8. Plano de Teste em Ambiente Escalado

Fase 1 - Load Testing (Semana 1-2):

- **Objetivo:** Validar limites atuais do sistema
- **Ferramentas:** JMeter, K6 para stress testing
- **Cenários:**
    - 100 usuários simultâneos

    - 1.000 uploads/hora sustentados
    - Picos de 2.000 requisições/minuto
- **Critérios de Sucesso:** Latência < 5s, Error rate < 1%

Fase 2 - Chaos Engineering (Semana 3):

- **Objetivo:** Testar resiliência do sistema
- **Ferramentas:** Chaos Monkey, Litmus
- **Cenários:**
    - Falha de nodes Kubernetes
    - Degradação de performance de GPU
    - Perda de conectividade de rede
- **Critérios de Sucesso:** Recovery automático < 2 minutos

Fase 3 - Production Simulation (Semana 4):

- **Objetivo:** Simular ambiente de produção real
- **Dados:** Dataset sintético com 10.000 imagens
- **Usuários:** 20 radiologistas simulados
- **Duração:** 48h contínuas
- **Métricas:** Performance, accuracy, resource usage

Fase 4 - Pilot Expansion (Mês 2):

- **Objetivo:** Teste com clientes reais (5 laboratórios)
- **Volume:** 500 exames/dia distribuídos
- **Monitoramento:** Real-time em todos os KPIs
- **Feedback Loop:** Weekly retrospectives
- **Go/No-Go Decision:** Baseado em métricas objetivas

Validação Contínua:

- **Canary Deployments:** 5% → 25% → 50% → 100%
- **A/B Testing:** Novas features testadas com subgrupo
- **Performance Regression Testing:** CI/CD pipeline
- **Disaster Recovery Drills:** Mensais

Conclusão

Este plano de escalabilidade garante evolução controlada do sistema de predição de idade óssea, mantendo qualidade clínica enquanto suporta crescimento exponencial. O foco em monitoramento proativo, testes rigorosos e mitigação de riscos assegura sucesso na transição de MVP para plataforma nacional de diagnóstico médico assistido por IA.

Next Steps:

1. Aprovação do budget de infraestrutura
2. Setup inicial do ambiente de staging escalado
3. Início da Fase 1 de testes de carga
4. Recrutamento de ML/DevOps engineers especializados