2. Documento: Implantação

1. Visão Geral

A fase de implantação visa operacionalizar o modelo treinado para que o **Mapa de Degradação de Habitats** se torne um produto utilizável pelos gestores do ecoturismo angolano. A estratégia é focar em um **Pipeline de Processamento Assíncrono na Nuvem** para lidar com a natureza intensiva em dados do Sensoriamento Remoto.

2. Serialização de Modelos

- **Formato de Serialização:** O modelo U-Net foi salvo no formato **H5** (unet_best.h5) usando a API Keras/TensorFlow.
- **Vantagens:** O formato H5 encapsula a arquitetura do modelo e os pesos otimizados em um único arquivo, garantindo a portabilidade e a facilidade de carregamento em qualquer ambiente Python com TensorFlow. Para ambientes de produção, uma alternativa seria o formato **TensorFlow SavedModel**, mais otimizado para *Model Serving*.
- **Armazenamento:** O arquivo .h5 seria armazenado em um serviço de *Object Storage* como o **Google Cloud Storage (GCS)**, garantindo durabilidade, alta disponibilidade e acesso de baixa latência para o serviço de inferência.

3. Modelo de Serviço

- **Estratégia: Pipeline de Inferência Assíncrona e Escalável**.
- **Justificativa:** A inferência em GeoTIFFs de satélite (imagens de alta resolução, multi-banda e de grande tamanho) é um processo que consome tempo e recursos computacionais. Uma API síncrona em tempo real é inadequada. A inferência é melhor executada em segundo plano.
- **Plataforma de Implantação: Google Cloud Platform (GCP)**.
    - **Início do Processo:** Um novo GeoTIFF de entrada é enviado ao **GCS**.
    - **Gatilho:** A chegada de um novo arquivo no GCS aciona um serviço *Serverless*, como o **Google Cloud Run** ou **Cloud Functions**.
    - **Execução da Inferência:** O serviço *Serverless* executa o pipeline de inferência (carregamento do .h5, *patching*, normalização, predição e reconstrução) em um ambiente isolado e escalável.
    - **Saída:** O GeoTIFF final (mapa_segmentacao_final.tif) é escrito de volta no GCS para ser acessado pelo usuário ou por um sistema GIS (Geographic Information System) externo.

4. Integração de API (Visão Futura: Sistema de Alerta)

Embora a inferência da cena completa não seja feita via API em tempo real, uma API seria crucial para um futuro **Sistema de Alerta de Degradação Precoce** para áreas menores.

- **Tecnologia: FastAPI** (Python) para criar um serviço web rápido.
- **Endpoint Proposto:** POST /api/v1/predict_patch
- **Formato de Entrada (JSON):**
- JSON

```
{
  "bounding_box": [lon_min, lat_min, lon_max, lat_max],
  "data_aquisicao": "YYYY-MM-DD"
}
```

- **Formato de Resposta (JSON/Binário):**
- JSON

```
{
  "status": "success",
  "mask_encoded": "base64_encoded_png_of_segmentation_mask",
  "main_class_detected": "Degradação"
}
```

5. Considerações de Segurança

- **Privacidade de Dados:** O projeto utiliza apenas dados de satélite públicos e gratuitos (Sentinel-2), o que minimiza os riscos de privacidade. Não são utilizados dados pessoais.
- **Autenticação e Autorização:** O acesso à **Cloud Run** e ao **GCS** será restrito apenas a utilizadores autorizados (por exemplo, gestores de conservação) através de políticas de **IAM (Identity and Access Management)** do GCP, garantindo que apenas entidades confiáveis possam acionar ou aceder aos resultados do modelo.

6. Monitoramento e Registro

Elemento Monitorado	Métrica	Ferramenta	Mecanismo de Alerta
Desempenho do Modelo	**Mean IoU** (monitorado periodicamente em novos dados).	Cloud Run Logs / Custom Metrics	Alerta acionado se o **Mean IoU cair abaixo de 0.60** (indicando **Model Drift**).
Infraestrutura	**Latência de Inferência** (Tempo total para processar um GeoTIFF).	Google Cloud Monitoring	Alerta se o tempo de processamento exceder um limite (ex: 30 minutos), indicando um problema de escala ou desempenho.
Integridade do Serviço	**Taxa de Erros HTTP** (Se a API futura falhar).	Google Cloud Logging	Alerta imediato se a taxa de erros for > 5% no período de 5 minutos.

O registro detalhado de cada execução de inferência (entrada, parâmetros, tempo de execução e classe predominante detectada) será feito no **Google Cloud Logging** para rastreabilidade e *troubleshooting*.