Caso de Uso Alternativo: Análise de Documentos Jurídicos com Pipeline em GCP + Open Source

Contexto

Um departamento jurídico de uma instituição pública precisa auditar milhares de documentos jurídicos digitalizados (contratos, petições, e-mails, pareceres) com foco em **identificação de riscos legais** e **extração de padrões de comportamento processual**. O objetivo é fornecer tanto:

- **Visão histórica completa** (batch), para entender padrões e atores recorrentes.
- **Análise em tempo real** (stream), para detectar imediatamente conteúdos críticos ou termos de risco.

Arquitetura Lambda com Tecnologias Open Source e Google Cloud

1. Fontes de Dados (Sensor Layer)

- Documentos: .pdf, .docx, .eml, .xlsx.
- Origem:
    - Upload manual via portal interno.
    - Integração com sistema de protocolo judicial.
    - Captura automática de e-mails com etiquetas jurídicas.

---

2. Ingestão e Armazenamento (Ingestion Layer)

- **Google Cloud Storage (GCS)** como Data Lake: gs://juridico-auditoria/raw-docs.
- Uploads de documentos disparam eventos via **Pub/Sub**, criando um sistema reativo.
- Cada documento é versionado e enriquecido com metadados (nome do autor, data, tipo).

---

3. Camada Batch (Batch Layer – Visão Histórica Completa)

- **Objetivo:** Processar diariamente todos os documentos do caso para gerar:
    - Índices de busca

    - Tópicos jurídicos recorrentes
    - Mapa de entidades e atores

Tecnologias Utilizadas:

- **Apache Beam** rodando no **Google Dataflow** para processamento em lote.
- **Tesseract OCR** (em container) para extrair texto de PDFs digitalizados.
- **spaCy** com modelo jurídico customizado para:
    - Reconhecimento de Entidades Nomeadas (NER)
    - Extração de cláusulas contratuais
- **BERTopic** para modelagem de tópicos (usando embeddings do Sentence-BERT)
- Resultados armazenados em:
    - **BigQuery** (para exploração via SQL e BI)
    - **ElasticSearch** (para busca textual avançada)

---

4. Camada Speed (Speed Layer – Análise em Tempo Real)

- Gatilho via **Pub/Sub** a cada novo documento.
- Função executada com **Cloud Functions** ou **Cloud Run**, com baixa latência.
- A pipeline em tempo real realiza:
    - OCR com Google Vision API
    - Classificação de urgência com modelo de ML no **Vertex AI**
    - Verificação de termos críticos (“quebra de sigilo”, “fraude contratual”, etc.)
- Alerta imediato enviado via **Firebase Cloud Messaging** ou Slack.

---

5. Serving Layer e Visualização

- Painel unificado desenvolvido com **Streamlit** (ou Dash) hospedado no **App Engine**.
- Camadas integradas via:

    - **ElasticSearch** para buscas
    - **BigQuery** para relatórios históricos e dashboards
- Recursos disponíveis:
    - Buscas avançadas por entidade, tipo de cláusula ou período
    - Mapa de comunicação entre partes jurídicas (via Neo4j)
    - Timeline dos documentos críticos identificados em tempo real

---

Exemplo do Fluxo

1. Equipe jurídica faz upload de 30.000 documentos para GCS.
2. O pipeline em lote executa à noite via Dataflow, processa tudo e atualiza os índices em BigQuery e ElasticSearch.
3. No dia seguinte, o advogado pesquisa "acordo sigiloso com fornecedor X" e encontra 18 ocorrências.
4. Durante o dia, um novo e-mail chega com a expressão “admissão de irregularidade”.
5. A Cloud Function detecta o termo de risco e dispara alerta em 5 segundos.
6. O gestor jurídico acessa o painel, lê o conteúdo e toma ação imediata.