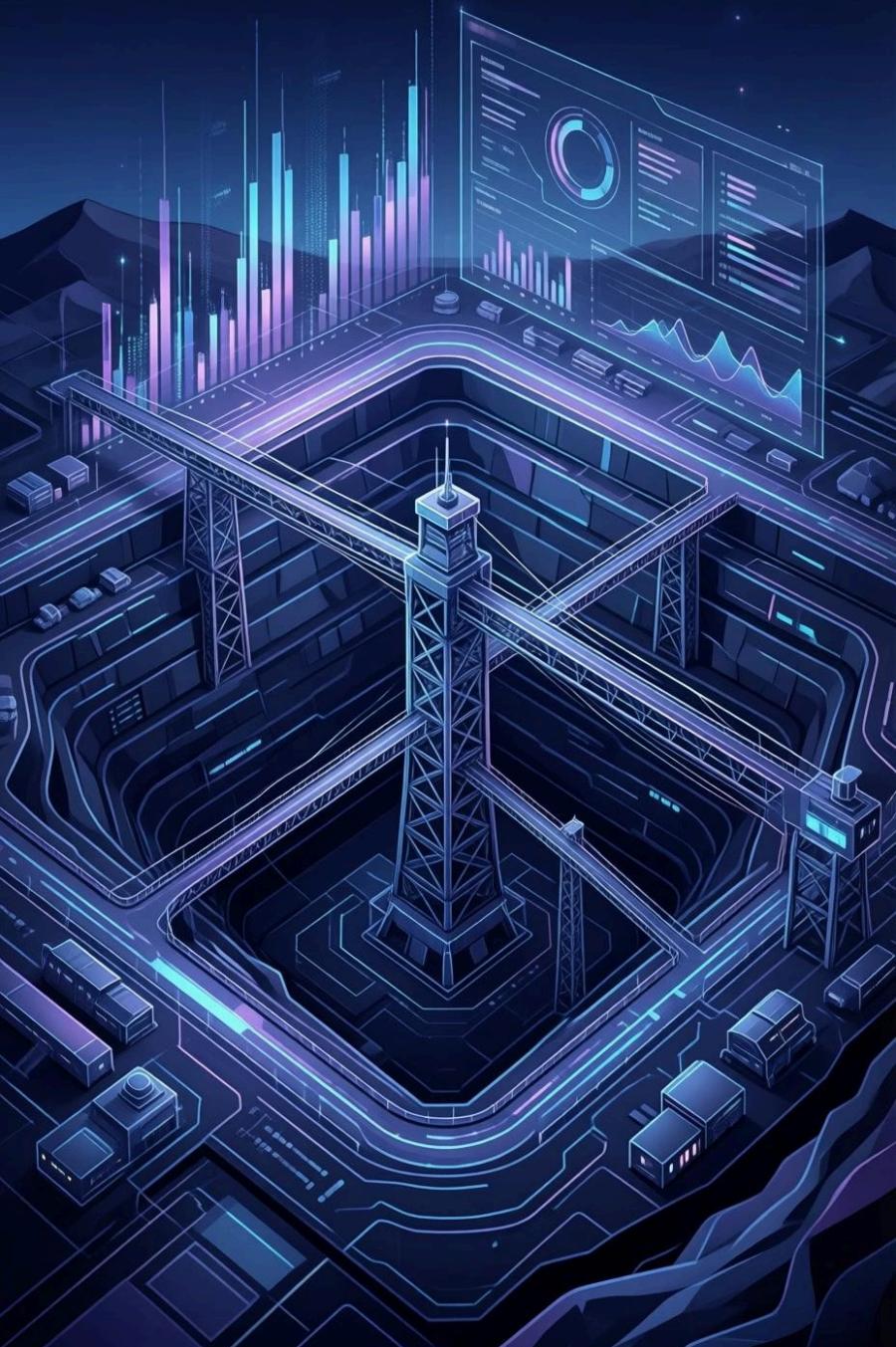


Ingesta y Estructuración Inteligente de Reportes Mineros

Pipeline Modular con LlamaIndex & GPT-4o-mini para la
automatización de extracción de datos NI 43-101

 ENERO 2026

 FRANKY - DATA ENGINEER



El Desafío de Negocio

Contexto del Sector

La industria minera genera gigabytes de reportes técnicos NI 43-101 en formato PDF, conteniendo datos críticos de inversión que permanecen no estructurados y difíciles de analizar a escala.



Documentos Extensos

Información dispersa en documentos de más de 200 páginas con estructura heterogénea y formatos inconsistentes.

Datos Atrapados

Tablas críticas embebidas en imágenes o formatos no extraíbles mediante técnicas tradicionales de parsing.

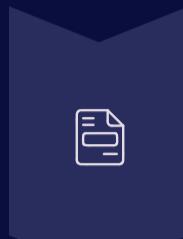
Ambigüedad Semántica

Terminología técnica variable: "Recursos Medidos" vs "Reservas Probadas" requiere interpretación contextual experta.

- ❑ **Misión:** Construir un pipeline inteligente capaz de transformar PDFs complejos en JSONs estrictamente tipados, listos para analítica avanzada y toma de decisiones de inversión.

Arquitectura de la Solución

Estrategia de **"Divide y Vencerás"** para maximizar la precisión del modelo de lenguaje y minimizar alucinaciones.



Ingesta Inteligente



Escaneo del PDF para mapear índices y estructura visual mediante LlamaIndex.



Extracción Modular



División en 4 fases especializadas: Metadata, Recursos, Reservas y Economía.



Validación Rigurosa



Pydantic fuerza esquemas JSON estrictos con reglas de negocio integradas.

Las 4 Fases de Extracción

01

Metadata del Proyecto

Identificación, ubicación geográfica y propietarios del yacimiento minero.

02

Recursos Minerales

Tablas de tonelaje, leyes de corte y categorización (Medidos, Indicados, Inferidos).

03

Reservas Mineras

Estimaciones económicas viables: Probadas y Probables con análisis de viabilidad.

04

Indicadores Económicos

CAPEX, OPEX, NPV, IRR y parámetros financieros clave del proyecto.



Stack Tecnológico

Selección orientada a **precisión** y **costo-eficiencia** en entornos de producción.



Llamaindex

Orquestación avanzada de contexto y retrieval inteligente de información estructurada.



GPT-4o-mini

Balance óptimo entre razonamiento complejo y bajo costo operativo (\$0.15/1M tokens).



Pydantic

Garantía de contrato de datos con validación automática y tipado estricto.



Python 3.10+

Ecosistema maduro con librerías especializadas para procesamiento de documentos.

¿Por qué GPT-4o-mini?

- Razonamiento sobre tablas complejas y contexto multi-página
- Reducción del 90% en costos vs. modelos grandes
- Latencia optimizada para procesamiento batch
- Soporte nativo para JSON estructurado



Resultados y Validación

El sistema procesa exitosamente documentos complejos, manejando errores comunes de OCR y minimizando alucinaciones del modelo.

100%

Documentos Procesados

Tasa de éxito en extracción de estructura básica

4

Fases Modulares

División estratégica del pipeline de extracción

JSON

Output Estandarizado

Formato mining_report_*.json tipado

Calidad de Datos Garantizada



Normalización de Unidades

Conversión automática entre sistemas métrico/imperial: toneladas, onzas troy, gramos por tonelada.



Manejo de Valores Nulos

Detección inteligente de celdas vacías, NA, y valores ausentes con estrategias de imputación definidas.



Sistema de Alertas

Clasificación automática Warning/Error integrada en el reporte con trazabilidad completa.

Performance Comprobado: Procesamiento robusto de múltiples archivos en modo batch con validación cruzada y detección de inconsistencias lógicas.

Propuesta de Producción

Cloud AWS NATIVE

Serverless

Arquitectura escalable para procesar **10,000+ PDFs** con alta disponibilidad y costos optimizados.

Ingesta (S3 + SQS)

Bucket raw-data con Event Notifications hacia cola SQS para procesamiento asíncrono.

Extracción (Lambda)

Funciones Python/LlamaIndex consumiendo colas SQS con timeout extendido.

1

2

3

4

OCR/Parsing (Batch)

Instancias GPU g4dn para conversión PDF→Markdown usando herramientas especializadas.

Almacenamiento Dual

DynamoDB para acceso transaccional + S3/Athena para analítica agregada.

Estrategia de Almacenamiento

DynamoDB

- JSONs completos indexados por Project ID
- Acceso rápido (<10ms) para consultas individuales
- TTL automático para datos históricos

S3 + Athena

- Data Lake para analítica cross-project
- Queries SQL sobre millones de registros
- Ejemplo: "Total de oro en yacimientos colombianos"

- Monitoreo:** CloudWatch Logs + X-Ray para trazabilidad de costos por documento, detección de cuellos de botella y alertas de SLA.

Conclusiones

1. Solución Técnicamente Probada

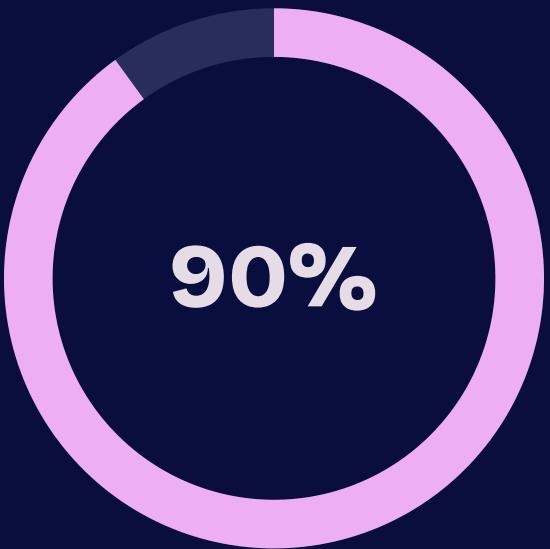
El script main.py demuestra la viabilidad completa de extraer datos financieros complejos de PDFs mineros, incluyendo manejo de tablas multi-página y validación de consistencia lógica.

2. Optimización de Costos

El uso estratégico de modelos "mini" optimizados y arquitectura de prompts modulares reduce el costo operativo en un **90%** comparado con modelos grandes, sin sacrificar precisión.

3. Preparado para Escala Empresarial

La arquitectura propuesta en AWS permite escalar de 5 documentos de prueba a 10,000+ reportes en producción sin fricción operativa, manteniendo latencias predecibles y costos lineales.



90%

Reducción de costos vs. modelos tradicionales



4

Fases modulares de extracción inteligente



10K+

Capacidad de documentos en producción

Gracias por su tiempo

Contacto

Franky Cardona

+57 3238821058

frankycardona1927@gmail.com

Próximos Pasos

- Piloto con 100 reportes reales
- Integración con sistemas BI existentes
- Expansión a otros estándares (JORC, SAMREC)

