

INFORME TÉCNICO DE VALIDACIÓN:

MÓDULO DE INGESTA OCR

PROYECTO: TrueStock

MÓDULO: Ingesta Automatizada de Documentos

SERVICIO: Amazon Textract

FECHA: 04 de Diciembre de 2025

RESPONSABLE: Equipo de Desarrollo (Backend/IA)

1. OBJETIVO DE LA PRUEBA

Validar la precisión, latencia y robustez del servicio de reconocimiento óptico de caracteres (OCR) implementado mediante **Amazon Textract** para la extracción automatizada de datos clave (Fecha, Proveedor, Total, Items) desde facturas físicas digitalizadas.

2. METODOLOGÍA DE PRUEBA

- **Dataset de Prueba:** Se recopiló un conjunto de datos de **60 facturas reales** de distintos proveedores chilenos.
- **Formatos de Entrada:** El dataset incluyó variaciones en formato PDF, JPG y PNG ² para simular condiciones reales de uso por parte de las PYMEs.
- **Entorno de Ejecución:** Las pruebas se ejecutaron en el entorno de producción (AWS), utilizando el flujo completo: *Frontend Next.js* → *API Gateway* → *S3 Textract* → *RabbitMQ*.
- **Criterio de Éxito:** Se considera "Éxito" cuando el sistema logra identificar y estructurar correctamente el 100% de los campos obligatorios (Nombre Proveedor, Fecha, Monto Total) y al menos el 90% del detalle de productos.

3. RESULTADOS OBTENIDOS

Métrica	Valor
Total de Documentos Procesados	60
Procesamiento Exitoso (Extracción Correcta)	58
Procesamiento Fallido / Incompleto	2
Tasa de Precisión (Accuracy)	96.67% (\approx 97%)
Tiempo Promedio de Respuesta	3.5 segundos por documento

4. ANÁLISIS DE FALLAS

Se analizaron los 2 casos fallidos (3% de error) para determinar la causa raíz:

1. **Caso 1 (Falso Negativo):** Factura capturada con muy baja luminosidad y ángulo excesivo.
2. **Caso 2 (Error de Formato):** Documento manuscrito ilegible donde el nivel de confianza de AWS fue insuficiente.

Acción Correctiva (Continuidad Operativa): El sistema implementa un diseño resiliente que no bloquea la operación ante fallos del OCR. Se dispone de una interfaz de Verificación y Edición donde el usuario puede revisar los datos extraídos o, en caso de fallo total, optar por el Ingreso Manual de los ítems, asegurando que el flujo de negocio (actualización de stock) nunca se detenga por problemas de imagen.

5. CONCLUSIÓN TÉCNICA

El módulo de OCR demuestra una **alta confiabilidad (97%)** en condiciones de operación normales. La arquitectura asíncrona basada en colas (RabbitMQ) permitió procesar el lote sin degradar el rendimiento de la aplicación web.

Se certifica el módulo como **APTO PARA PRODUCCIÓN**, cumpliendo con los requerimientos funcionales de automatización de entrada de datos definidos en el alcance del proyecto.