

VNA-SB-020 | Oracle Data, AI & Industry Accelerators | AI Document Processing & Oracle 23ai Vector Search | Redwood Stone

Oracle 23ai Document AI & Intelligent OCR: Extract PDFs & Invoices into APEX & Vector Search

Analysts and operations clerks waste thousands of hours manually reading scanned PDFs, supplier invoices, master service agreements (MSAs/SOWs), and regulatory filings to key fields into ERP and database tables. Our Document AI pipeline combines OCI Document Understanding, vision-language models, and Oracle Database 23ai AI Vector Search with a human-in-the-loop Oracle APEX 24.2 exception workbench.

Search Queries & Scope Covered: Oracle 23ai AI Vector Search DBMS_VECTOR RAG pipeline | OCI Document Understanding invoice OCR to Oracle APEX | human in the loop PDF verification workbench Oracle APEX | extract unstructured PDF contracts to Oracle JSON Duality Views | automate 3-way PO invoice matching in Oracle Database

SOURCE LEGACY STACK (STACK X)		MODERNIZED TARGET ARCHITECTURE (STACK Y)	
Unstructured PDFs, Scanned TIFFs, Invoices, MSAs/SOWs & Lab Reports		OCI Document Understanding, Multimodal LLMs, Oracle 23ai AI Vector Search & APEX	
FIELD EXTRACTION ACCURACY	MANUAL KEYING TIME CUT	VECTOR & RELATIONAL STORE	EXCEPTION REVIEW UI
96-99.4% Verified	85%+ Reduction	Single Oracle 23ai DB	Side-by-Side APEX Viewer

1. Problem Space & Architectural Bottlenecks

- Legacy template-based OCR tools break the moment a vendor changes their invoice layout or a multi-page contract contains nested tables and handwritten annotations.
- Connecting standalone third-party AI tools to enterprise ERPs requires stitching together separate vector databases (Pinecone/Weaviate), blob stores, Python glue servers, and custom review UIs.
- Regulated finance, healthcare, and legal workflows require deterministic confidence scoring and a fast side-by-side human verification screen whenever a field confidence score falls below threshold.

2. Technical Solution & Reference Architecture

- 2.1 Multi-Stage Ingestion & Layout-Aware OCR:** Incoming documents from email inboxes, SFTP, or OCI Object Storage are processed through OCI Document Understanding / Veryfi / Nanonets and vision-language models using strict JSON Schema output enforcement.
- 2.2 Unified Oracle 23ai Relational + JSON + AI Vector Storage:** Instead of maintaining a separate vector database, extracted JSON payloads, `VECTOR` embeddings (`DBMS\_VECTOR`), and relational ledger rows live inside a single ACID-compliant Oracle Database 23ai instance.
- 2.3 Automated Entity Resolution & PO/Contract Matching:** Extracted line items are automatically cross-checked against existing purchase orders, contract rate cards, and master vendor tables using hybrid SQL + Vector similarity queries.
- 2.4 Human-in-the-Loop APEX 24.2 Verification Workbench:** Documents with confidence scores below 95% or math discrepancies route to a side-by-side APEX split-screen viewer highlighting the exact bounding box on the PDF next to the editable form field.

3. Artifact-by-Artifact Conversion & Modernization Matrix

Legacy Artifact / Baseline Workload	Modernized Oracle / Cloud Target	Engineering Method & Outcome
Manual Email / PDF Invoice & Receipt Entry	Automated OCR Pipeline + 3-Way PO Match in Oracle 23ai	Straight-Through Processing + Bounding-Box Exception UI
Unindexed PDF Contract / MSA / SOW Repositories	Structured Renewal/Rate Tables + Oracle 23ai AI Vector Search	Clause Extraction + In-Database RAG (`DBMS_VECTOR`)
Siloed External Vector DBs & Python Glue Apps	Converged Oracle 23ai (Relational + JSON Duality + Vector)	Single Security, Backup & SQL Governance Plane
Private Equity / Financial PDF Trial Balance Packs	Automated Medallion Lakehouse Ingestion & Validation	Deterministic Cross-Foot & Balance Tie-Out Rules

4. Oracle APEX & Ancillary Engineering Toolchain

- Oracle Database 23ai (`VECTOR` data type, `DBMS\_VECTOR`, JSON Relational Duality Views)
- OCI Document Understanding, OCI Generative AI (Cohere / Llama), and Oracle Functions `F(n)`
- Oracle APEX 24.2 Split-Screen PDF Bounding-Box Verification Workbench
- Oracle REST Data Services (ORDS) for webhook ingestion

Anonymized Production Proof: From agricultural genetics and retail accounts-payable automation to private-equity portfolio reporting and global import/customs manifest tracking, our Document AI + Oracle 23ai pipelines process ratusan of thousands of unstructured documents directly into governed enterprise schemas.

5. Principal Architect Q&A

Q: Why store vector embeddings inside Oracle Database 23ai instead of a dedicated vector database?

A: In enterprise applications, you almost never run a pure vector search in isolation; you need to filter by customer ID, effective date, department security role, and contract status at the same time. Oracle 23ai executes hybrid relational + JSON + vector similarity joins in a single SQL query with zero data duplication.

Q: How do you prevent LLM hallucinations on financial numbers and dates?

A: Every extracted document passes through deterministic PL/SQL validation rules (line-item sum == invoice subtotal + tax == total, valid PO lookup, date range bounds). Any mathematical mismatch is flagged automatically in the APEX human-in-the-loop queue.

Q: What is included in the downloadable PDF Solution Brief?

A: The PDF Solution Brief includes the Oracle 23ai Converged Document AI Reference Architecture, sample `DBMS\_VECTOR` hybrid SQL patterns, and our human-in-the-loop confidence routing matrix.