

Amazon Textract – Short Note **Definition:**

Amazon Textract is a fully managed machine learning (ML) service by AWS that automatically extracts text, handwriting, and data from scanned documents, PDFs, and images.

Key Features:

- **Text Extraction:** Detects printed text and handwriting from documents.
- **Form Data Extraction:** Identifies key-value pairs in forms (e.g., Name → John Doe).
- **Table Extraction:** Extracts tabular data accurately while preserving structure.
- **Integration:** Easily integrates with AWS services like S3, Lambda, and Comprehend.
- **OCR + AI:** Goes beyond traditional OCR by understanding document layout and context.

Use Cases:

- Invoice and receipt processing
- Automated form data entry
- Identity document analysis
- Document digitization and archiving

Advantages:

- High accuracy using ML models
- Scalable and serverless
- Secure (integrates with AWS IAM and encryption)