

ドキュメント正規化ツール：企業データとAIをつなぐ架け橋

非構造化ドキュメントを、AIが活用可能な「資産」へ変換する



課題：AIは社内ドキュメントをそのまま読めない

ChatGPTやCopilot導入の最大の障壁は、データの「非構造化」にあります。



形式がバラバラ

PDF、Word、Excel、PowerPointなど、社内には統一されていない形式のファイルが散乱しています。



画像内のテキスト

スキャンされたPDFや写真として貼りけられた文字情報は、通常のテキスト抽出では読み取れません。



レイアウト情報の喪失

単純なテキスト抽出を行うと、表の構造や段組みが崩れ、意味が通じないデータになってしまいます。



メタデータの欠落

ファイル名、ページ番号、作成日時などの重要なコンテキスト情報が、変換過程で失われがちです。

解決策：正規化による「構造化データ」への変換

画像、表、メタデータを保持したまま、AIが理解しやすいJSON形式に統一します。

Input / 入力



非構造化データ
(Unstructured)

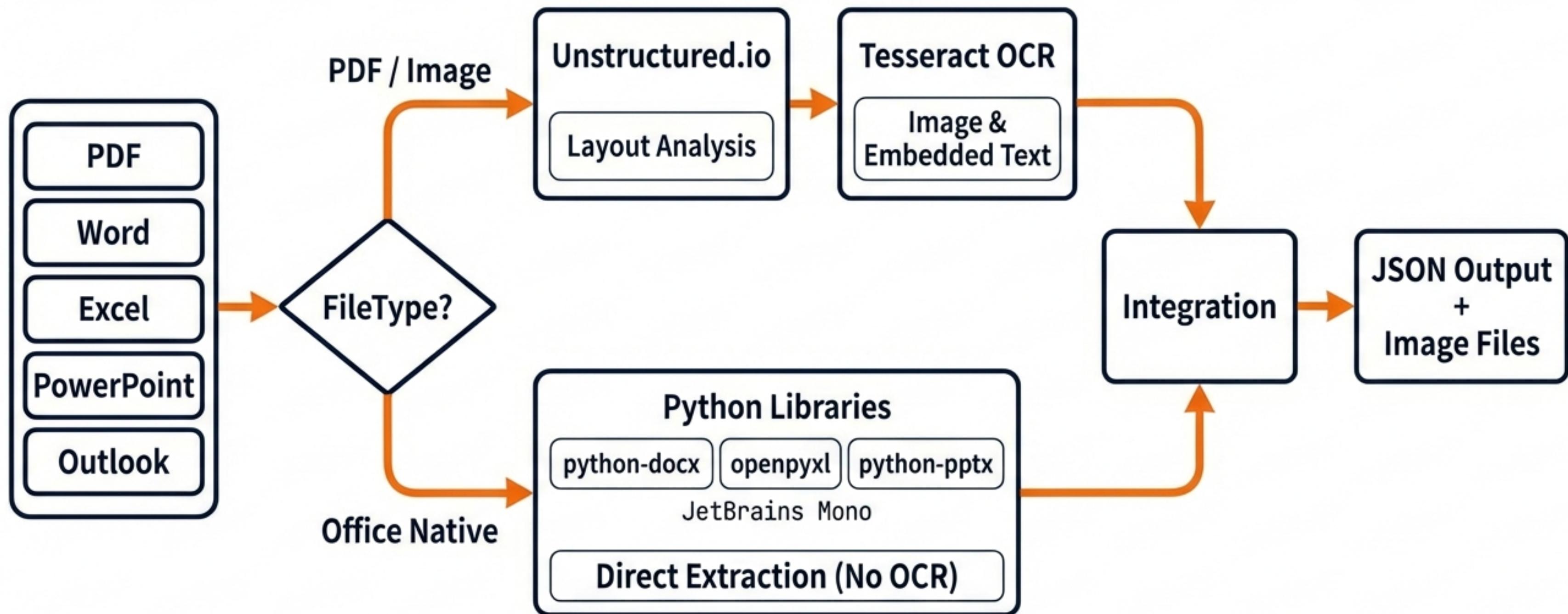


正規化エンジン
(Normalization Engine)

Output / 出力



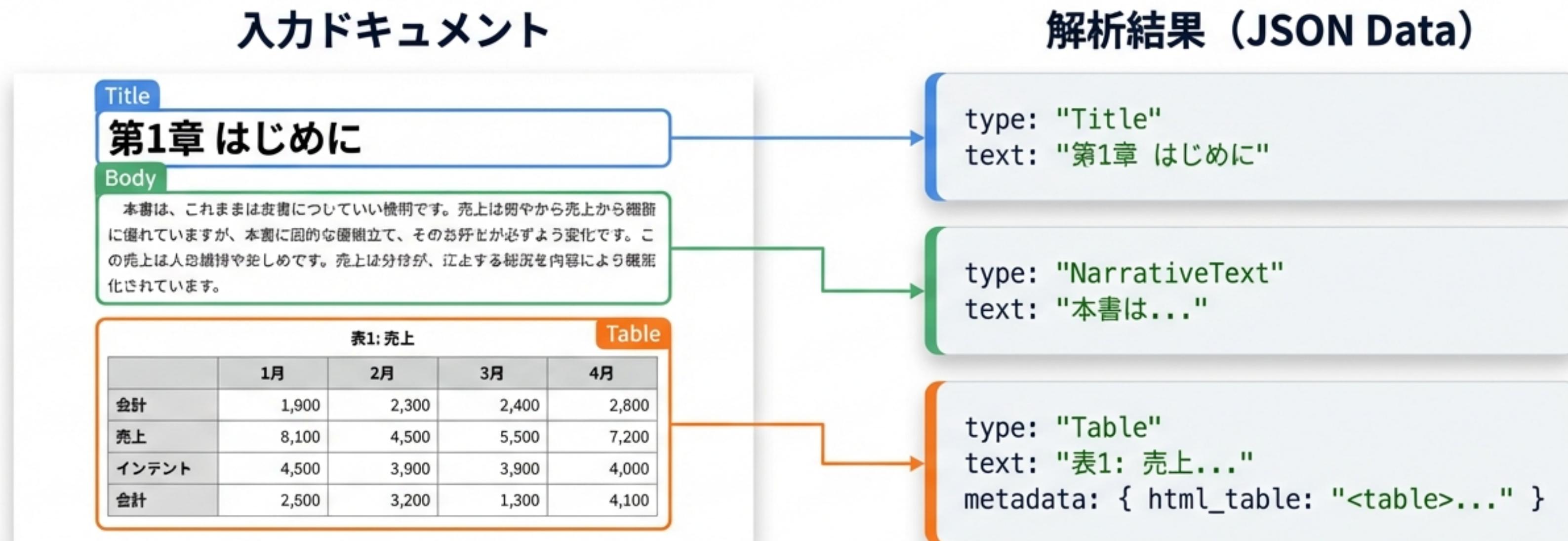
システムアーキテクチャ概要



ファイル形式に応じて最適な解析エンジンを自動選択します。

ドキュメント解析エンジン：Unstructured.io

hi_res（高解像度）モードによる要素識別



- 要素タイプ識別 (Title, Body, Table, Image)
- 表の構造保持 (HTMLテーブルとして保存)

OCR (光学文字認識) : 画像内の情報をデータ化



Technology: Tesseract OCR

- 対応言語: 日本語 (jpn) + 英語 (eng)
- 処理対象: スキャンPDF, 埋め込み画像, グラフ・図表内のテキスト

```
extracted_text = "第1四半期の売上は前年比120%を達成..."
```

Office形式のネイティブ解析

OCRを経由しないため、100%正確なテキスト抽出とスタイル情報の保持が可能です。

Target Extension	Python Library	Extracted Elements
 .docx	python-docx	テキスト, 表, 画像, スタイル (Heading, Bold, etc.)
 .pptx	python-pptx	スライド, テキストボックス, 画像, ノート (Speaker Notes)
 .xlsx	openpyxl	セル値, 数式, シート構造

Native Advantage:

- 1. No OCR Errors
- 2. Fast Processing
- 3. Metadata Retention

出力データの構造：JSONアナトミー

```
{  
  "source_file": "2024年度報告書.pdf",  
  "total_chunks": 45,  
  "chunks": [  
    {  
      "chunk_id": "chunk_001",  
      "text": "第1章 エグゼクティブサマリー",  
      "element_type": "Title",  
      "page_number": 1,  
      "metadata": { "languages": ["jpn"] }  
    },  
    ...  
  ]  
}
```

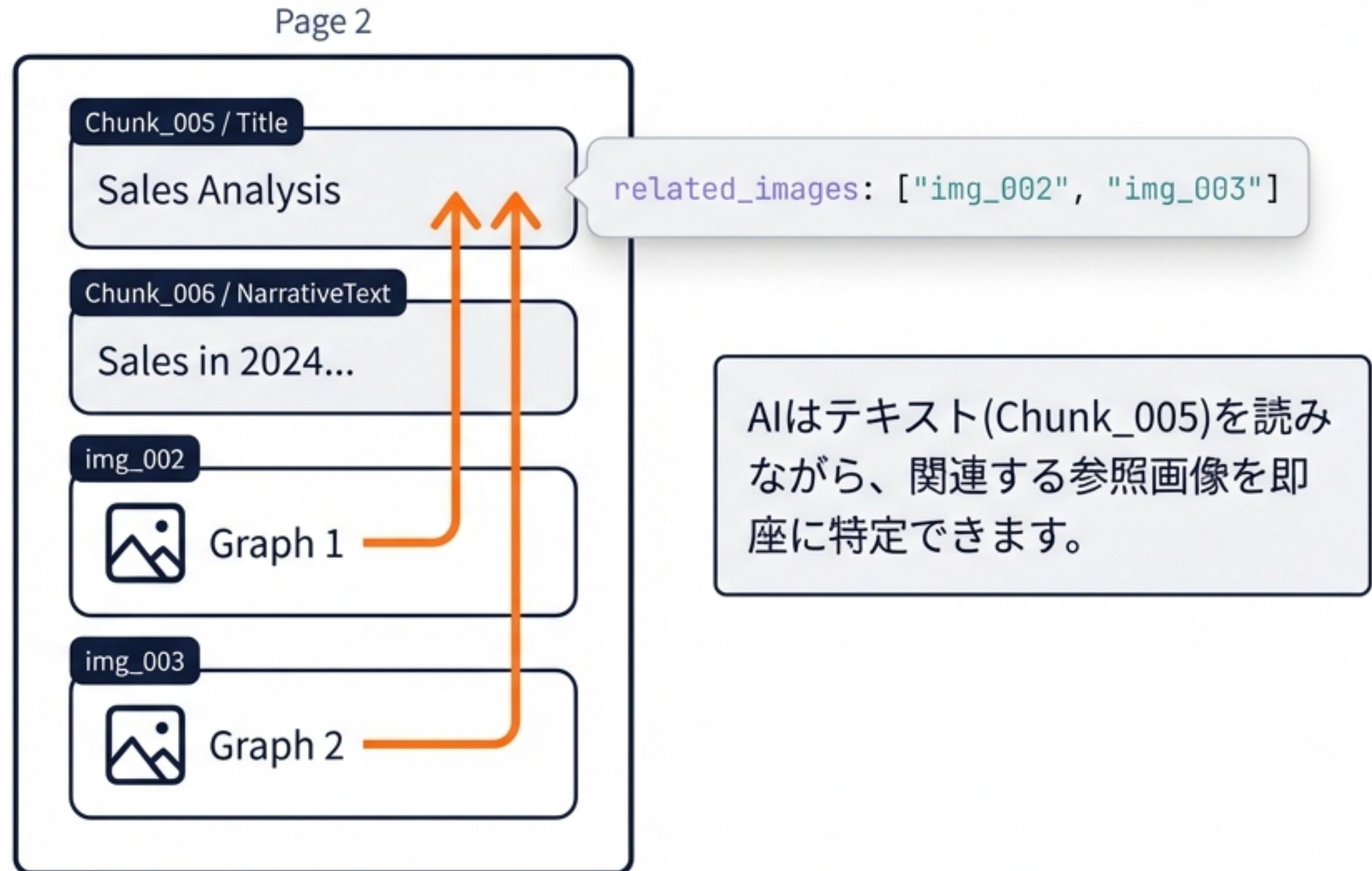
Chunks: 文書を意味のある
ブロック単位に分割

Element Type: テキストの
役割（タイトル、本文、表
など）を分類

Metadata: ページ番号や言
語情報を保持

画像とテキストの紐付けロジック

同じページにある画像は、そのページの「最初のチャンク」に紐付けられます。



データ検証と構造化：Pydantic

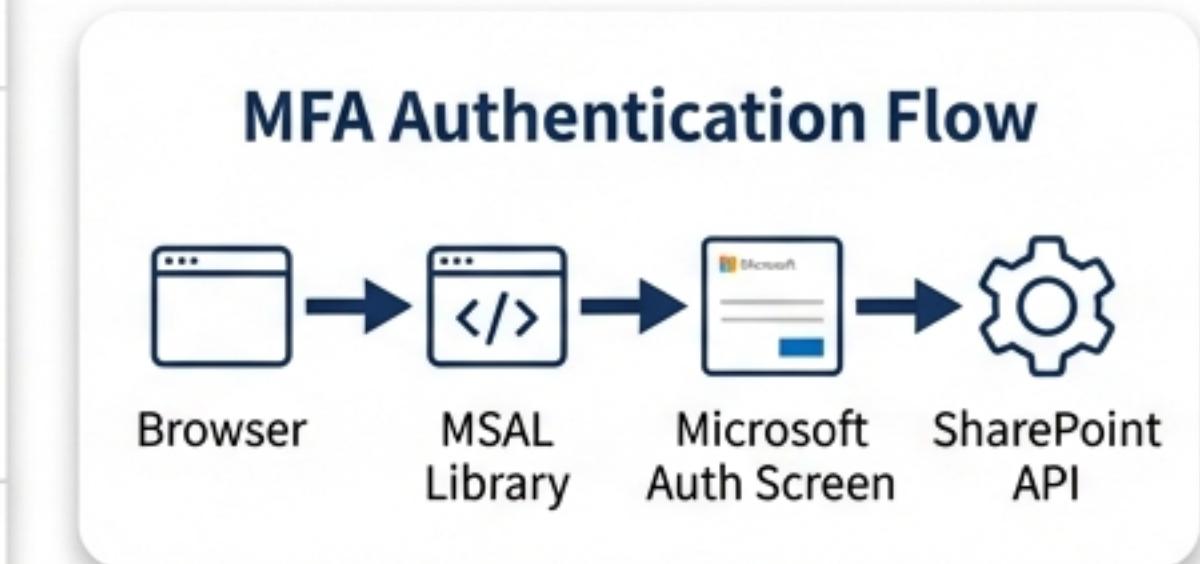
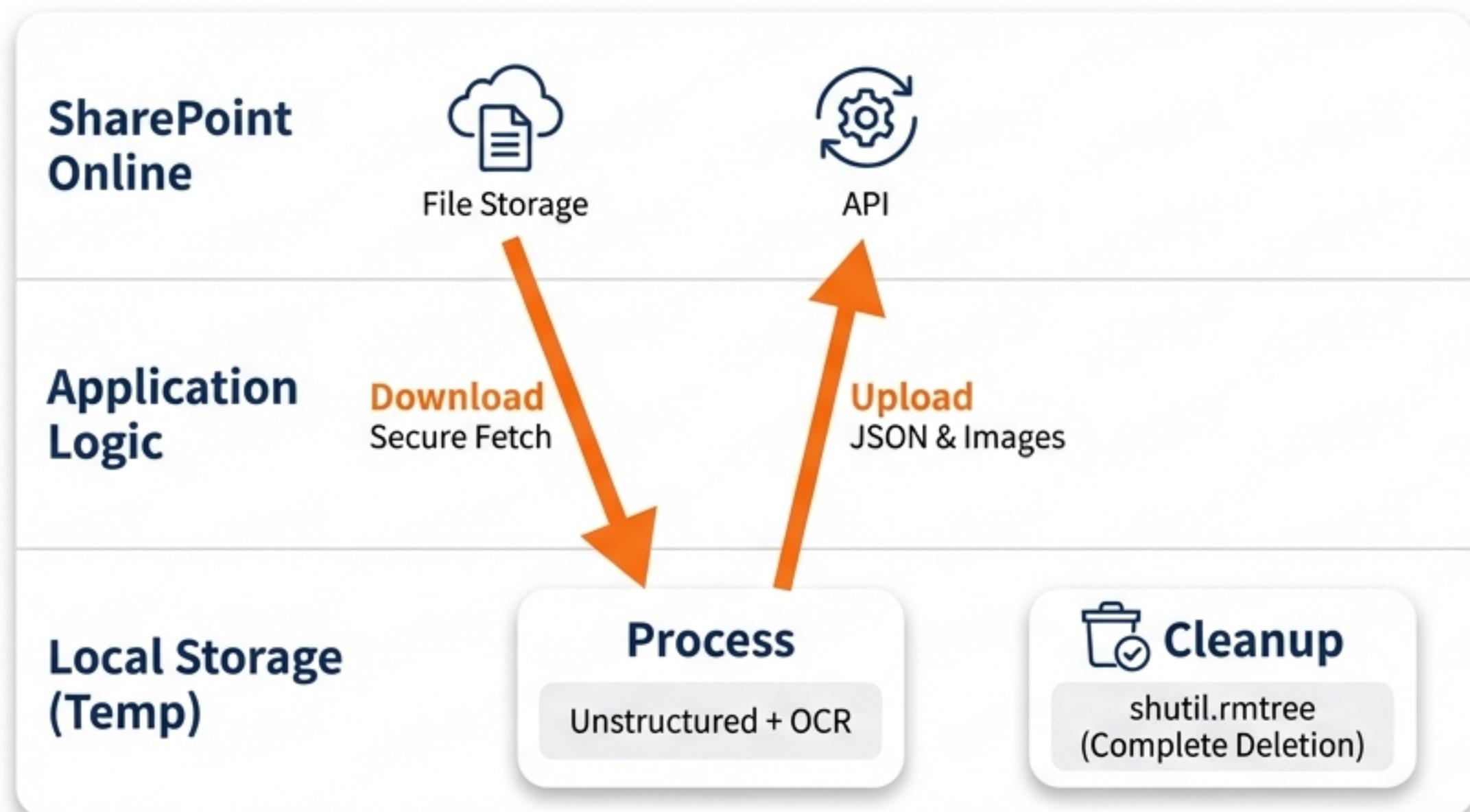
厳格な型定義により、出力JSONの整合性を常に保証します。

- Type Safety
- Error Handling
- Consistency

```
class NormalizedDocument(BaseModel):  
    source_file: str  
    total_chunks: int  
    total_images: int  
    chunks: List[Chunk]  
  
class Chunk(BaseModel):  
    chunk_id: str  
    text: str  
    element_type: str  
    page_number: int  
    related_images: List[str]
```



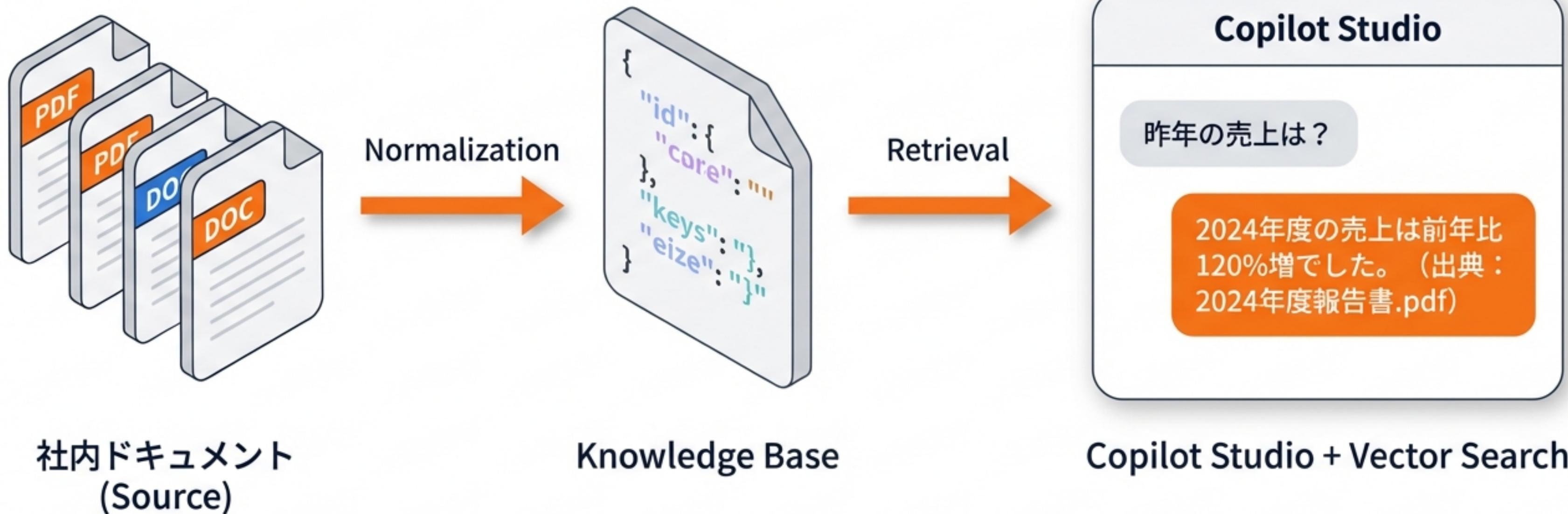
SharePoint連携とセキュリティ



一時ファイルは処理完了後に即座に完全削除され、ローカルストレージを圧迫しません。

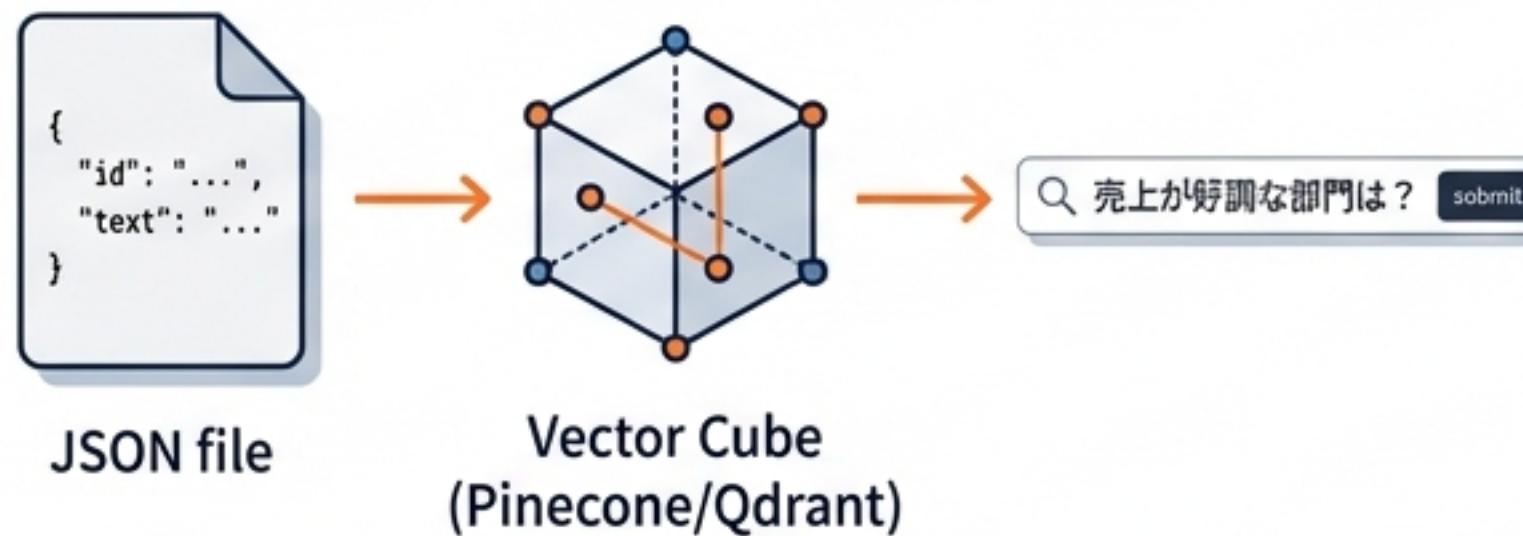
活用シーン1：社内AIチャットボット

正規化されたJSONをナレッジソースとして登録し、回答精度を向上



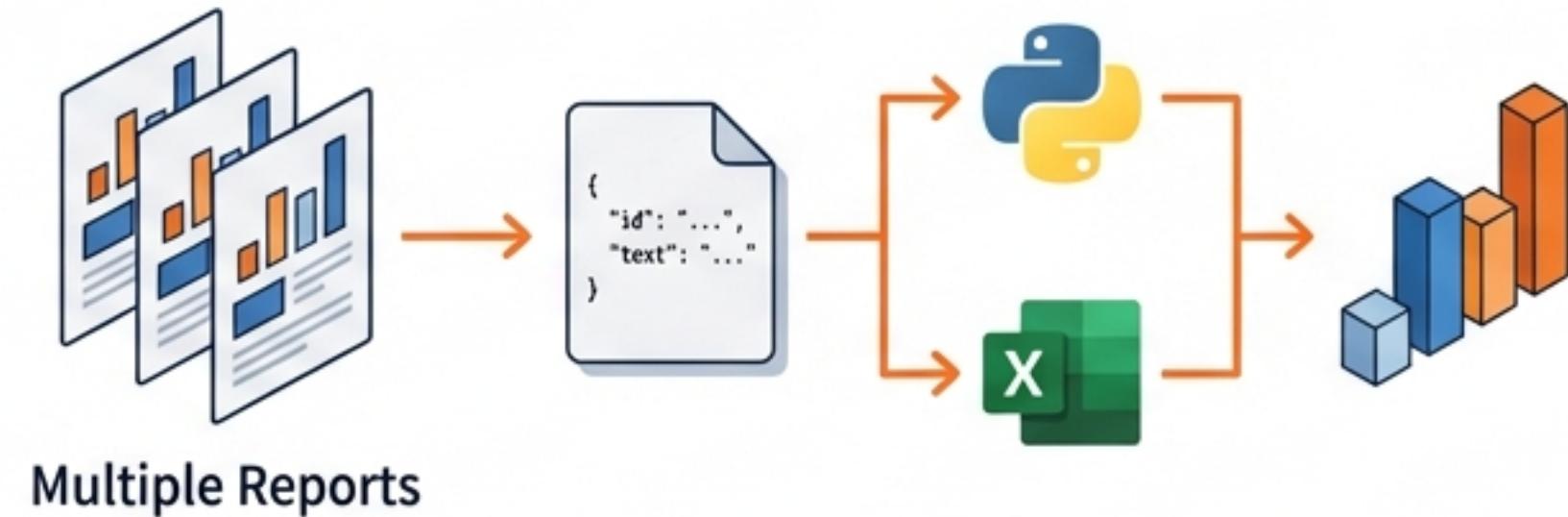
活用シーン 2 & 3：検索システムとデータ分析

Semantic Search System



キーワードだけでなく、意味に基づいた検索が可能に。
(例：「売上が好調な部門は？」)

Automated Analytics



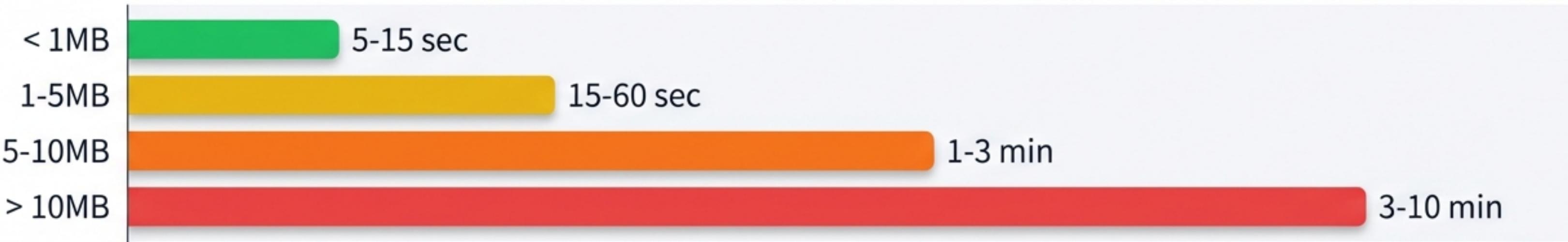
Multiple Reports

大量の報告書から数値を自動抽出し、集計・分析を自動化。
(例：全支店の月次データを統合)

構造化データ化により、検索や集計の自動化が可能になります。

パフォーマンスと技術スタック

Processing Time Estimate



Tech Stack Versions

unstructured (v0.16+)

pytesseract (v0.3+)

pillow (v10.0+)

pydantic (v2.0+)

※OCR処理（スキャンPDF）を含む場合、処理時間が増加します。

まとめ：DXを加速するデータ基盤

- ✓ **Multi-Format:** PDF, Word, Excel, PowerPoint, Outlook 全てに対応
- ✓ **OCR Integrated:** 画像やスキャンデータ内の文字も抽出
- ✓ **Structure Preserved:** 表やレイアウトの情報を保持
- ✓ **Metadata Rich:** ページ番号やファイル情報を記録
- ✓ **Secure Integration:** SharePoint連携 & MFA対応

このツールは、眠っている社内ドキュメントを
「AIが活用できる知識」へと変えます。