Week 7 | 课时 2 | 智能文档解析：布局、层级、表格、页码、坐标与 provenance

Table of contents

Parser 的价值不在“抽得多”，而在“保留得准”

这一讲进入 Week7 的解析层：

一个可用 parser 不只是输出文本，还要输出 layout、hierarchy、tables、page、bbox、provenance 和 parser capability。

进入课时 3 返回 Week7 总览

下载讲义

提供适合离线阅读的 PDF 版和适合批注整理的 Word 版。

PDF 版 · 打印 / 离线阅读 Word 版 · 批注 / 二次整理

这节课解决什么问题

到了解析层，最容易犯的错误是做工具横评：

> Docling、Unstructured、Azure Document Intelligence 到底哪个抽文本效果更好?

这不是 Week7 的核心问题。Week7 真正要问的是：

哪个路线能在 Student Core 里稳定保留结构 metadata，并能被 contract、test 和 quality gate 验证？

参考学习时间

45–55 分钟

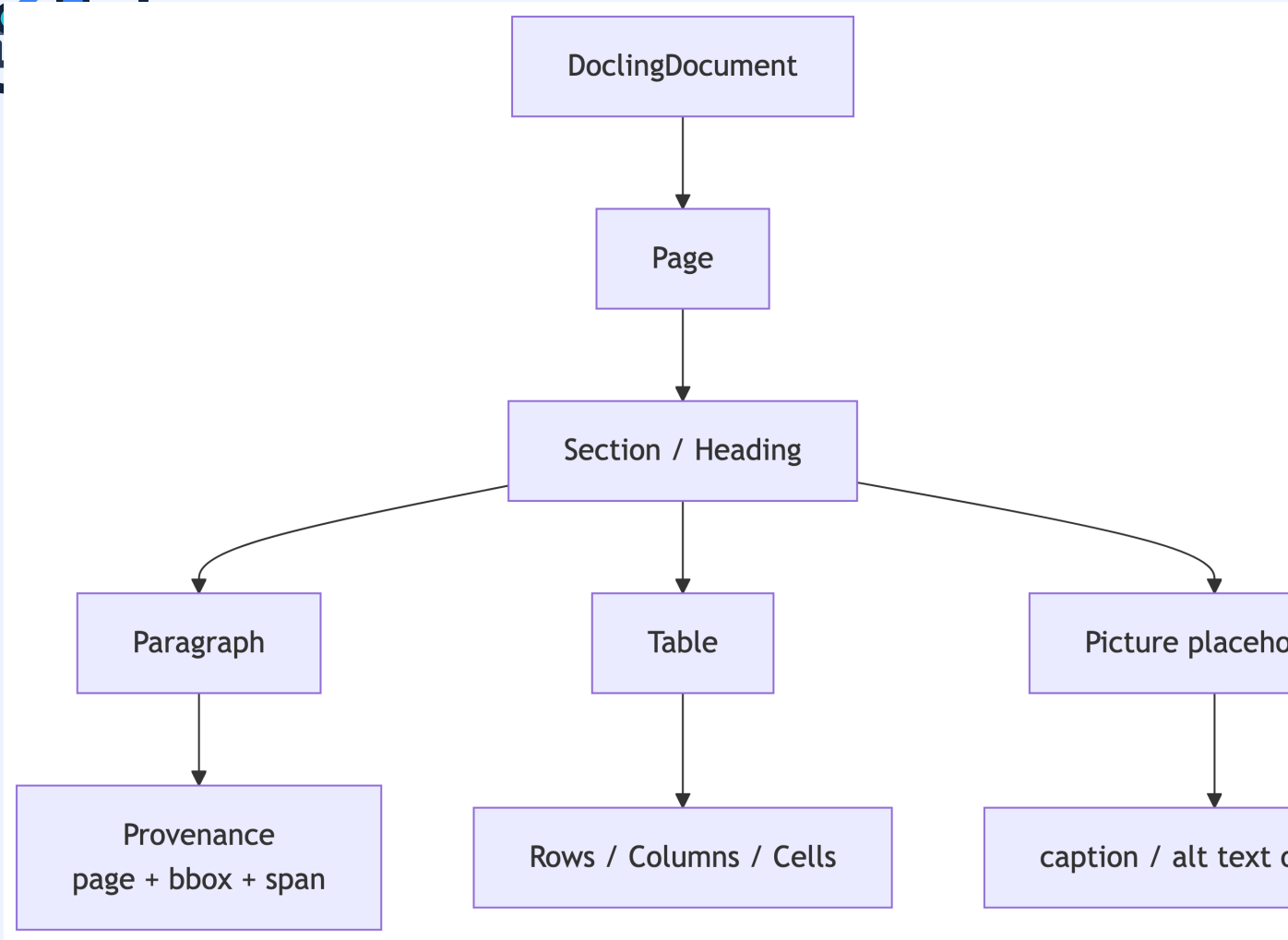
学完这一讲，你应该能做到什么

1. 说明 Docling-first 为什么适合 Week7 Student Core。
2. 判断 parser 输出是否保留了 page、bbox、section path 和 provenance。
3. 区分 parser capability、parser output 和 quality gate。
4. 说明为什么 OCR、ASR、video、VLM caption 不作为本周必需路径。

本课产出

- docs/blueprints/week07/adr-parser-adapter-route.md
- contracts/data/knowledge_section.schema.json
- contracts/data/parse_run.schema.json

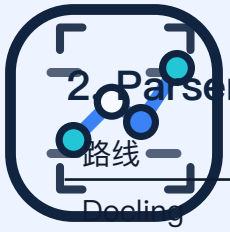
DoclingDocument 层级图



1. Student Core 为什么 Docling-first

Docling 路线适合本周，不是因为它“永远效果最好”，而是因为它天然贴近 Week7 的对象模型：

- 能表达 document hierarchy
- 能保留 page / bbox / provenance
- 能输出统一 document representation
- 能把 table、picture、text 分成不同结构对象
- 能本地执行，避免 Student Core 被云服务凭证和费用卡住



2. Parser adapter 决策表

路线	本周角色	适合做什么	不适合做什么
Docling	Student Core 默认	本地解析、结构保真、provenance、table / picture placeholder	不保证所有扫描件 OCR 都完美
Unstructured	optional fallback	elements 与 chunking 对照、某些格式 fallback	不作为第一必需路径
Azure Document Intelligence	Instructor optional	layout / table / bounding polygon 强演示	不适合作为学员必需依赖
OCR / ASR / video	Scale Pack	复杂 PDF、音频、视频扩展演示	不进入 Student Core 默认路径

3. Parser 输出最少要归一成什么

```
{
  "section_id": "stable_sha256_prefix",
  "doc_id": "workspace_help_center_v1",
  "section_type": "paragraph",
  "section_path": ["Setup", "SSO", "Troubleshooting"],
  "page_no": 4,
  "bbox": {"x0": 0.13, "y0": 0.22, "x1": 0.84, "y1": 0.31},
  "bbox_missing_reason": null,
  "content": "If SSO callback fails, verify the redirect URI...",
  "source_fingerprint": "sha256...",
  "doc_version": "v1",
  "parse_strategy_version": "docling_v1_no_ocr"
}
```

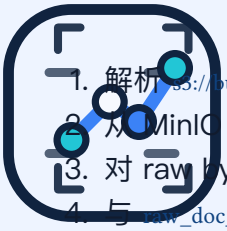
这个 JSON 不是展示用字段，而是 Week8 citation、质量门禁和策略回归的事实基础。

4. MinIO raw object 是主路径

Week3 的 doc_ingest.py 会把文档写到类似这样的路径：

```
s3://omni-raw-documents/workspace-helpcenter/source.pdf
```

Week7 不能只从本地目录读文件。正确顺序是：



1. 解析 s3://bucket/key
2. 从 MinIO 下载 raw bytes
3. 对 raw bytes 重新计算 source_fingerprint
4. 与 raw_doc_asset.source_fingerprint 比对
5. 指纹不一致时 hard fail 或 quarantine
6. MinIO 失败时才允许 source_dir fallback，并写入 fallback_reason

本课最重要判断

Parser 不是工具选择题，而是 **结构元数据的生成器**。选型标准是：输出能否支撑 contract、quality gate 和 Week8 citation。

自检清单

- 我能解释 Docling–first 的 Student Core 理由。
- 我知道 Unstructured 和 Azure DI 的边界。
- 我能列出 parser 输出必须保留的字段。
- 我知道 source_fingerprint 必须基于真实 raw bytes 校验。

最小行动命令

```
python -m pipelines.parse_normalize.parsers.docling_adapter \
  --input s3://omni-raw-documents/workspace-helpcenter/source.pdf \
  --out artifacts/week07/parsed_doc.json \
  --no-ocr
```