Week 7｜课时 1｜为什么“抽出文本”不等于非结构化数据工程

Table of contents

不是把 PDF 变成文本，而是把文档变成可追溯的数据资产

这一讲先立住 Week7 的核心判断：

非结构化数据工程的产出不是一堆 text chunks，而是能被检索、引用、质检和回放的 document asset。

进入课时 2 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

Week03 已经能把文档拿进系统，但这还不等于 Week08 可以安全检索。

如果解析阶段只留下纯文本，下游会立刻遇到这些问题：

- 表格被打散，字段和含义断开

- 标题层级丢失，chunk 不知道自己属于哪个 section
- 页码和 bbox 丢失，citation 无法回指原文
- 解析策略变化后，坏案例无法复盘
- 质量不合格的 chunk 被直接送进索引

所以 Week7 的目标是把 raw document 升级成：

parsed document + knowledge sections + document chunks + evidence anchors + quality gate。

参考学习时间

45–55 分钟

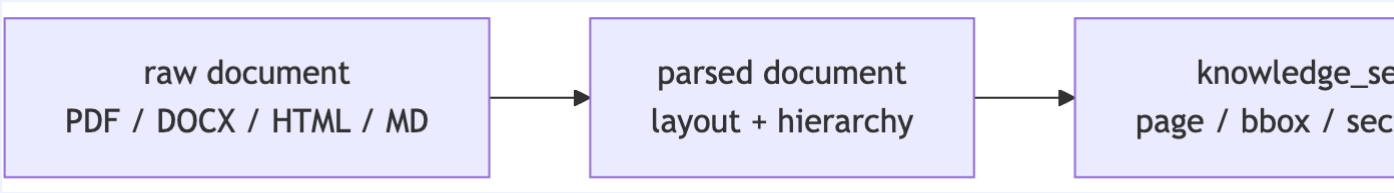
学完这一讲，你应该能做到什么

1. 区分 raw document、parsed document、section、chunk、anchor。
2. 说明为什么“PDF to text”不能直接等价于 RAG 可用输入。
3. 解释 Week7 和 Week8 的边界。
4. 写出 Week7 document asset v1 的最小字段。
5. 判断一个解析结果是否具备进入 Week8 的基本资格。

本课产出

- docs/blueprints/week07/week07-execution-blueprint.md
- docs/blueprints/week07/document-asset-boundary.md

先看一张总图



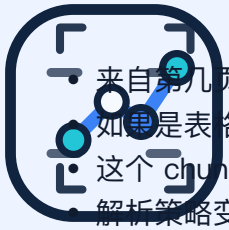
1. “抽出文本”为什么不够

很多 RAG demo 会把文档处理简化成：

```
PDF -> text -> chunks -> embeddings
```

这条链路最大的问题不是“效果不够好”，而是工程对象不完整。它没有回答：

- 这段文字来自哪个文件版本?



- 来自第几页、哪个标题层级?
- 如果是表格，表头和行列关系还在不在?
- 这个 chunk 是否有证据锚点?
- 解析策略变更后，坏案例能不能复现?

> ! 核心判断
>
> **非结构化数据工程不是抽文本，而是把文档变成可追溯、可质检、可版本化、可供检索消费的数据资产。**

2. Week7 的对象边界

对象	它是什么	不能混淆成什么
raw document	原始文件或对象存储路径	已可检索内容
parsed document	带布局和层级的解析结果	一串无结构文本
knowledge section	文档结构单元	最终检索 chunk
document chunk	Week8 索引候选单元	生成阶段的 citation
evidence anchor	回指原文位置的证据对象	LLM 临时拼接的 URL
quality gate	Week8 准入判断	人工随便看几个样本

3. Week7 和 Week8 的分界线

Week7 只做索引前的数据资产准备：

- parse
- normalize sections
- chunking
- evidence anchors
- quality report
- Week8 ready gate

Week8 才做：

- embedding
- pgvector / search index
- hybrid retrieval
- reranker
- RAG API

- citation rendering

如果 Week7 抢跑 Week8，会让课程主线变乱：学生还没有可验收 chunk，就开始讨论检索效果。

4. 一个合格 document asset 最少保留什么

字段	为什么必须保留
source_fingerprint	确认解析的是哪一份原始字节
doc_version	支撑版本回放和策略回归
page_no	PDF citation 的最小条件
bbox / bbox_missing_reason	支撑位置回指或解释为什么没有 bbox
section_path	保留标题层级和上下文
parse_strategy_version	解析策略可复现
chunk_strategy_version	chunk 策略可回归
quality_flags	告诉 Week8 哪些内容可消费、哪些要隔离

本课最重要判断

Week7 的输出不是“更多文本”，而是 **带结构、证据和质量状态的文档资产**。

自检清单

- 我能解释 raw document 和 parsed document 的区别。
- 我能说清为什么 citation 不能由 LLM 临时生成。
- 我知道 Week7 不做 embedding、检索和 RAG API。
- 我能列出 document asset v1 的关键字段。

最小行动命令

```
# 先确认 Week3 文档输入是否存在
python -m pipelines.ingestion.doc_ingest \
  --manifest data/seed_manifests/manifest_workspace_helpcenter_v1.json \
  --source-dir data/canonization/documents \
  --batch-id week07-lesson01-smoke \
  --report-json reports/week07/doc_ingest_input_report.json
```