



جاودان



انفجار داده‌ها

وقتی یک سیستم نرم‌افزاری شروع به کار می‌کند، به مرور حجم بسیار زیادی داده تولید می‌شود؛ کاربران ثبت‌نام می‌کنند، سفارش ثبت می‌کنند، فایل آپلود می‌کنند، و لاگ‌ها ذخیره می‌شوند.



داده‌ها وجود دارند، اما قابل استفاده نیستند.

این داده‌ها معمولاً در یک مکان و با یک فرمت نیستند. آنها در سیستم‌های مختلف پراکنده شده‌اند.

دیتابیس‌ها

فایل‌ها

APIهای مختلف

سرویس‌های ابری



مهندسی داده (Data Engineering)

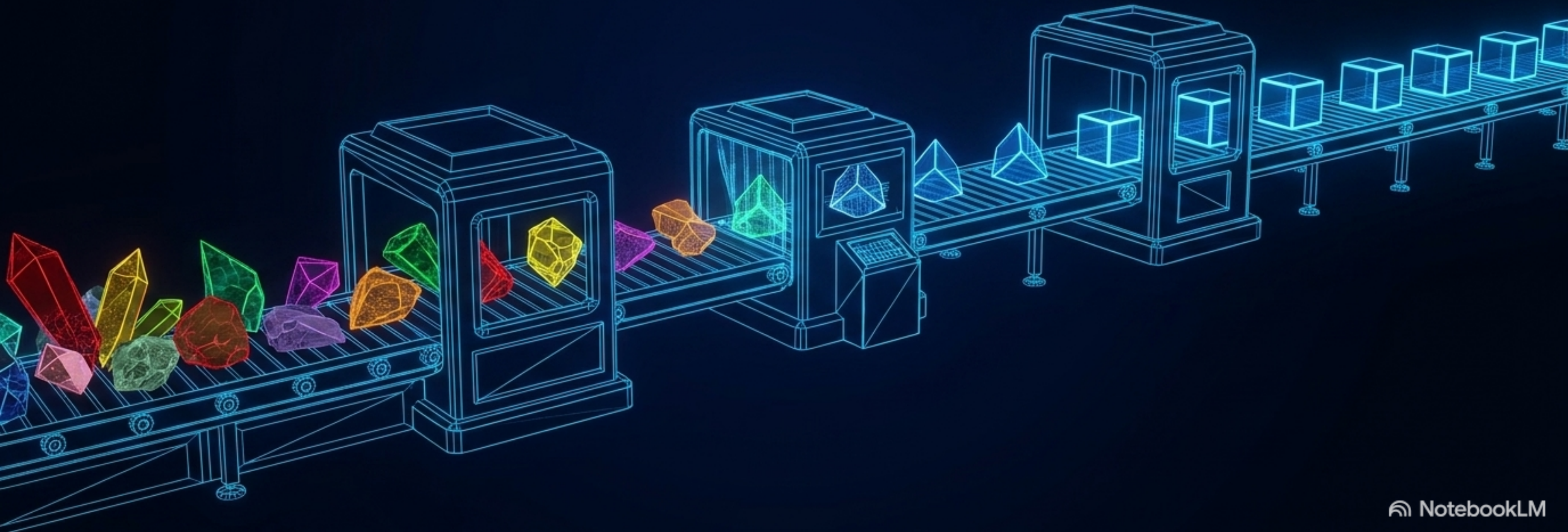
مجموعه‌ای از روش‌ها و ابزارهاست که این داده‌های پراکنده را جمع‌آوری، پردازش، تمیز و آماده می‌کند.

بدون مهندسی داده، داده وجود دارد اما قابل استفاده نیست.



مفهوم Data Pipeline

پایپ‌لاین داده را می‌توان یک خط تولید کارخانه تصور کرد. مواد خام وارد می‌شوند، پردازش و استاندارد می‌شوند، و در نهایت به شکلی درمی‌آیند که نرم‌افزار یا مدل هوش مصنوعی بتواند از آن استفاده کند.





GARBAGE IN → GARBAGE OUT

اگر داده‌ها را مستقیماً وارد برنامه کنیم، سیستم تحلیلی یا AI با شکست مواجه می‌شود.

Warning State

داده‌های تکراری ⚠

اطلاعات ناقص ⚠

سرعت پایین ⚠

افزایش خطا ⚠

Clean State

کارهای خودکار و استاندارد ✓



اجزای اصلی یک Data Pipeline

هر ابزاری که در مهندسی داده وجود دارد معمولاً یکی از این چهار مرحله را انجام می‌دهد.



مرحله ۱: Collect (جمع‌آوری)

اولین مرحله دریافت اطلاعات است. هدف این مرحله فقط جمع‌آوری داده است؛ هنوز هیچ پردازشی انجام نمی‌شود.





مرحله ۲: Process (پردازش)

بعد از دریافت داده، باید آن را اصلاح کرد.
در این مرحله کیفیت داده بالا می‌رود.



✓ حذف داده‌های تکراری

✓ اصلاح فرمت تاریخ

✓ حذف اطلاعات خراب

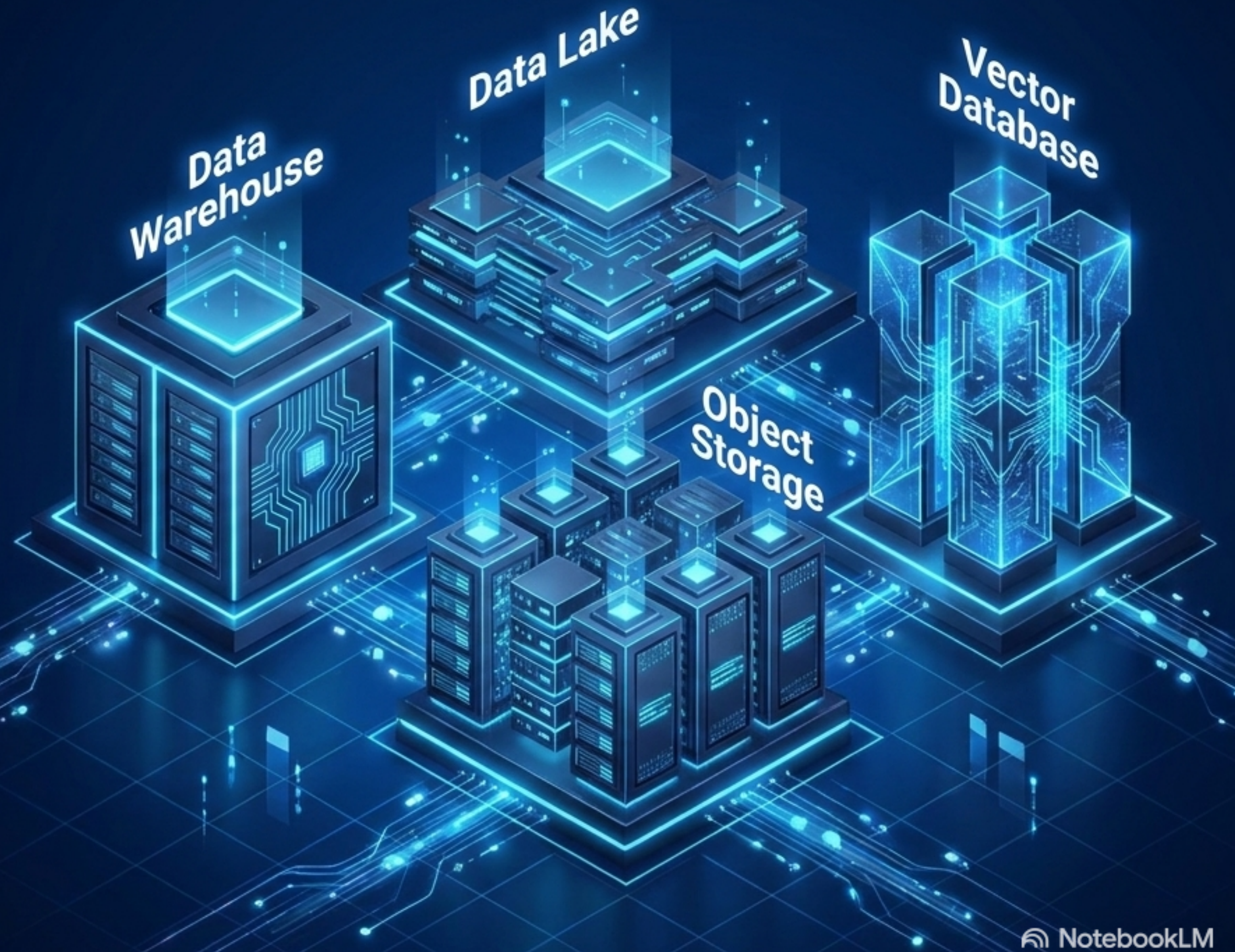
✓ نرمال‌سازی متن

✓ اعتبارسنجی داده



مرحله ۳: Store (ذخیره)

بعد از پردازش، داده باید ذخیره شود.
ذخیره‌سازی درست باعث می‌شود مراحل
بعدی سریع‌تر انجام شوند.





مرحله ۴: مصرف (Consume)

حالا داده آماده استفاده است. هدف
نهایی پایپلاین، رساندن داده آماده به
این بخش است.

Node 4



داشبورد مدیریتی



مدل
Machine
Learning



مدل LLM



RAG /
موتور جستجو

مسیر سنتی: ETL

(Extract → Transform → Load)

ابتدا داده پردازش می‌شود و بعد وارد دیتابیس نهایی می‌شود. این روش سال‌ها استاندارد اصلی صنعت بوده است.





تغییر پارادایم: از ETL به ELT

امروزه اکثر پروژه‌های بزرگ به دلیل حجم بالای داده‌ها و سرویس‌های ابری، به سمت ELT حرکت کرده‌اند.

ETL (کلاسیک)

- پردازش قبل از ذخیره
- مناسب داده کمتر
- کنترل بیشتر

ELT (مدرن)

- ذخیره قبل از پردازش
- مناسب Big Data
- مناسب Cloud
- سرعت بیشتر



پایپ‌لاین برای هوش مصنوعی (AI Pipelines)

در پروژه‌های هوش مصنوعی، پایپ‌لاین فقط داده عددی جابه‌جا نمی‌کند؛ بلکه فایل‌های متنی (PDF، Word، ایمیل، صفحات وب) را نیز پردازش می‌کند تا اطلاعات برای مدل‌های زبانی آماده شود.





مفهوم Embedding (تعبیه)

مدل‌های هوش مصنوعی متن را مانند انسان درک نمی‌کنند. آن‌ها متن را به مجموعه‌ای از اعداد تبدیل می‌کنند. به این نمایش عددی از متن، Embedding گفته می‌شود.

فکر کنید که شما یک متن را می‌خوانید و می‌خواهید آن را به یک مدل هوش مصنوعی ارائه دهید. مدل هوش مصنوعی نمی‌تواند متن را درک کند، اما می‌تواند اعداد را درک کند. بنابراین، شما متن را به یک سری اعداد تبدیل می‌کنید. این اعداد، Embedding نام دارند. این اعداد، مدل هوش مصنوعی را قادر می‌سازند تا متن را درک کند و به آن پاسخ دهد.

مثلاً، اگر شما متن "پادشاه" را به یک مدل هوش مصنوعی ارائه دهید، مدل هوش مصنوعی می‌تواند به شما پاسخ دهد که "پادشاه" چیست. این پاسخ، به دلیل این است که مدل هوش مصنوعی، Embedding متن "پادشاه" را درک کرده است.

Embedding ها را می‌توان به صورت یک ماتریس نمایش داد. در اینجا، یک ماتریس 4x4 از اعداد نمایش داده شده است:

0.124	-0.892	0.3252	0.8223
0.311	-0.342	0.2368	0.111
0.174	-0.892	-0.3010	0.124
0.882	0.111	0.8792	0.882
0.888	-0.111	-0.3262	0.124
0.836	-0.342	0.2492	0.882
0.896	-0.342	-0.3193	0.073
0.392	-0.111	-0.3866	0.111
0.127	-0.111	-0.3863	0.115
		-0.2894	0.832

درک شباهت معنایی (Semantic Similarity)

کلمات کاملاً متفاوت هستند، اما مفهوم مشابه است.
جمله‌هایی که مفهوم مشابه دارند، Embedding
نزدیک‌تری در فضای برداری خواهند داشت.





مسیر تولید Embedding

بعد از آماده شدن متن، این ۴ مرحله طی می‌شود.
از این لحظه داده برای جستجوی معنایی آماده
آماده است.

1

تقسیم متن

2

ارسال به مدل

3

تولید بردار عددی

4

ذخیره در دیتابیس



داده آماده برای RAG

مدل‌های زبانی نباید مستقیماً فایل PDF چندصد صفحه‌ای را بخوانند. پایپ‌لاین ابتدا متن را استخراج می‌کند، آن را Chunk (بخش‌بندی) می‌کند، Embedding می‌سازد و ذخیره می‌کند. این فرآیند RAG Ready Data تولید می‌کند.



معماری نهایی RAG

کیفیت و طراحی صحیح Pipeline تأثیر مستقیمی بر سرعت، دقت و عملکرد نهایی سیستم‌های هوش مصنوعی دارد. این زنجیره، پایه و اساس سیستم‌های داده‌محور است.

