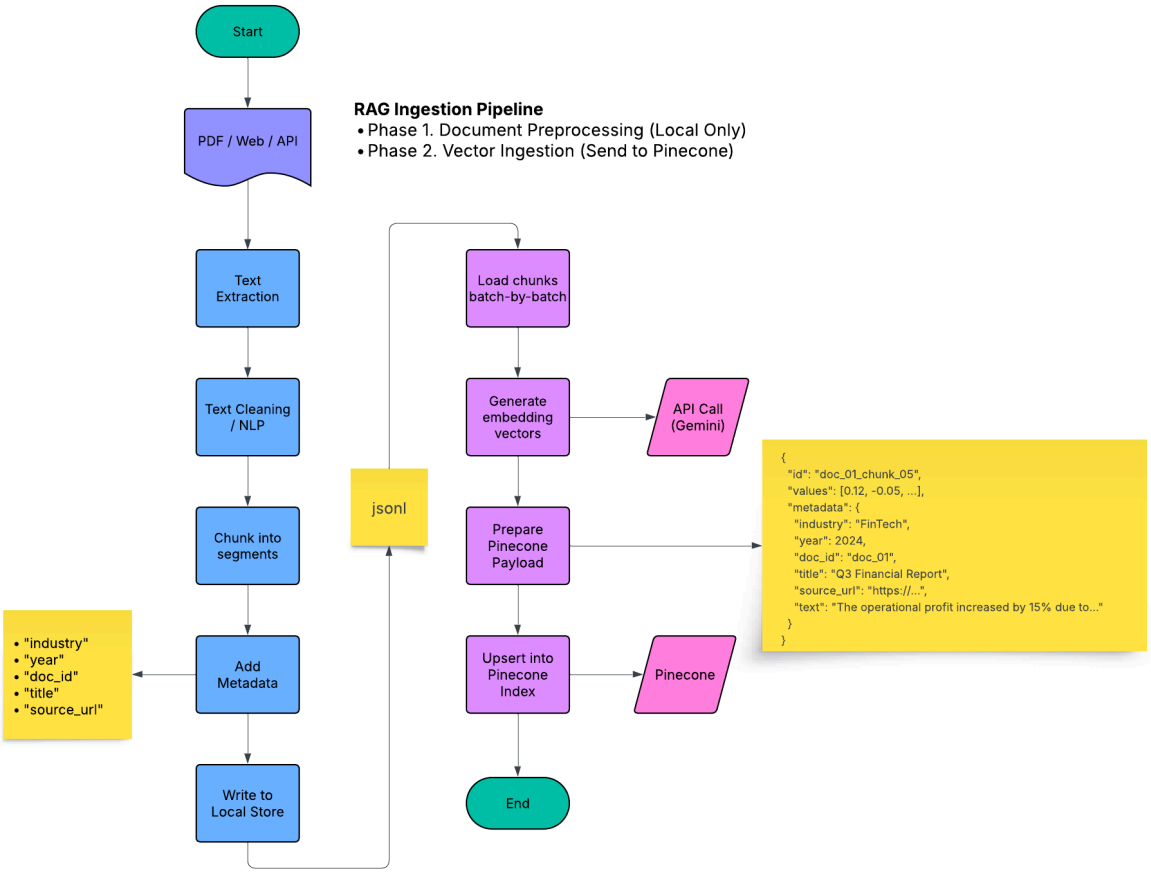
RAG Ingestion Pipeline-diagram



模拟数据集为MIT AI News Published till 2023

加载方式如下

```
# Install dependencies as needed:
# pip install kagglehub[pandas-datasets]
import kagglehub
from kagglehub import KaggleDatasetAdapter
```

```
# Set the path to the file you'd like to load
file_path = ""

# Load the latest version
df = kagglehub.load_dataset(
  KaggleDatasetAdapter.PANDAS,
  "deepanshudalal09/mit-ai-news-published-till-2023",
  file_path,
  # Provide any additional arguments like
  # sql_query or pandas_kwargs. See the
  # documenation for more information:
  # https://github.com/Kaggle/kagglehub/blob/main/README.md#kaggledatas
etadapterpandas
)

print("First 5 records:", df.head())
```

推荐的单个 JSONL 格式如下

```
{
  "id": "doc_01_chunk_05",
  "text": "The operational profit increased by 15% due to...",
  "metadata": {
    "industry": "FinTech",
    "year": 2024,
    "doc_id": "doc_01",
    "title": "Q3 Financial Report",
    "source_url": "https://...",
  }
}
```

存入pinecone内部的格式如下

```
{
  "id": "doc_01_chunk_05",
  "values": [0.12, -0.05, ...], // Gemini 生成的向量
  "metadata": {
    // --- 用于过滤 (Filtering) ---
    "industry": "FinTech",        // 必须：用于区分行业上下文
    "year": 2024,                 // 必须：用于按年份筛选 (Int类型方便比大小)

    // --- 用于前端展示 (Citation) ---
    "doc_id": "doc_01",           // 用于去重或分组
    "title": "Q3 Financial Report", // 用于前端显示的引用标题
    "source_url": "https://...",  // 用于点击

    // --- 核心内容 (Context) ---
    "text": "The operational profit increased by 15% due to..." // ⚠️ 必须：RAG 召回的实际文本
  }
}
```