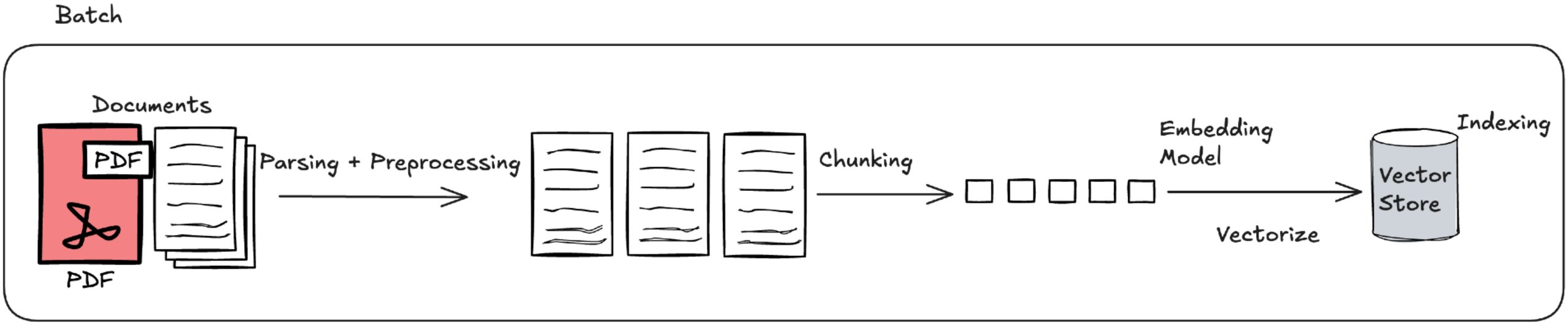
Retrieval Augmented Generation(검색 증강 생성)

- Foundation 모델은 일반적인 목적을 가지고 만들어졌기 때문에 대부분의 경우에 사용자의 컨텍스트를 이해할 수 없다.
- RAG는 응답을 생성하기 전에 신뢰할 수 있는 지식 베이스를 참조하도록 하는 프로세스
- LLM에 질문을 하기 전에 미리 정리해둔 데이터베이스를 참조해서 유저의 질문에 대한 컨텍스트를 만들고, 그 컨텍스트를 사용하여 유저의 질문과 함께 LLM에게 전달



Embedding Model

"Ollama is the easiest way to get up and running with large language models."

```
[-0.1552158921546936,
 -0.3130679428577423,
 -0.2622824013233185,
 -0.1073082387447357 2,
 ...
 0.2600640952587 1277,
 0.1449477970600 1282,
 -0.0151495300233364 1,
 0.0440374761819839 5]
```

https://ollama.com/blog/embedding-models