Clase 8 – RAG

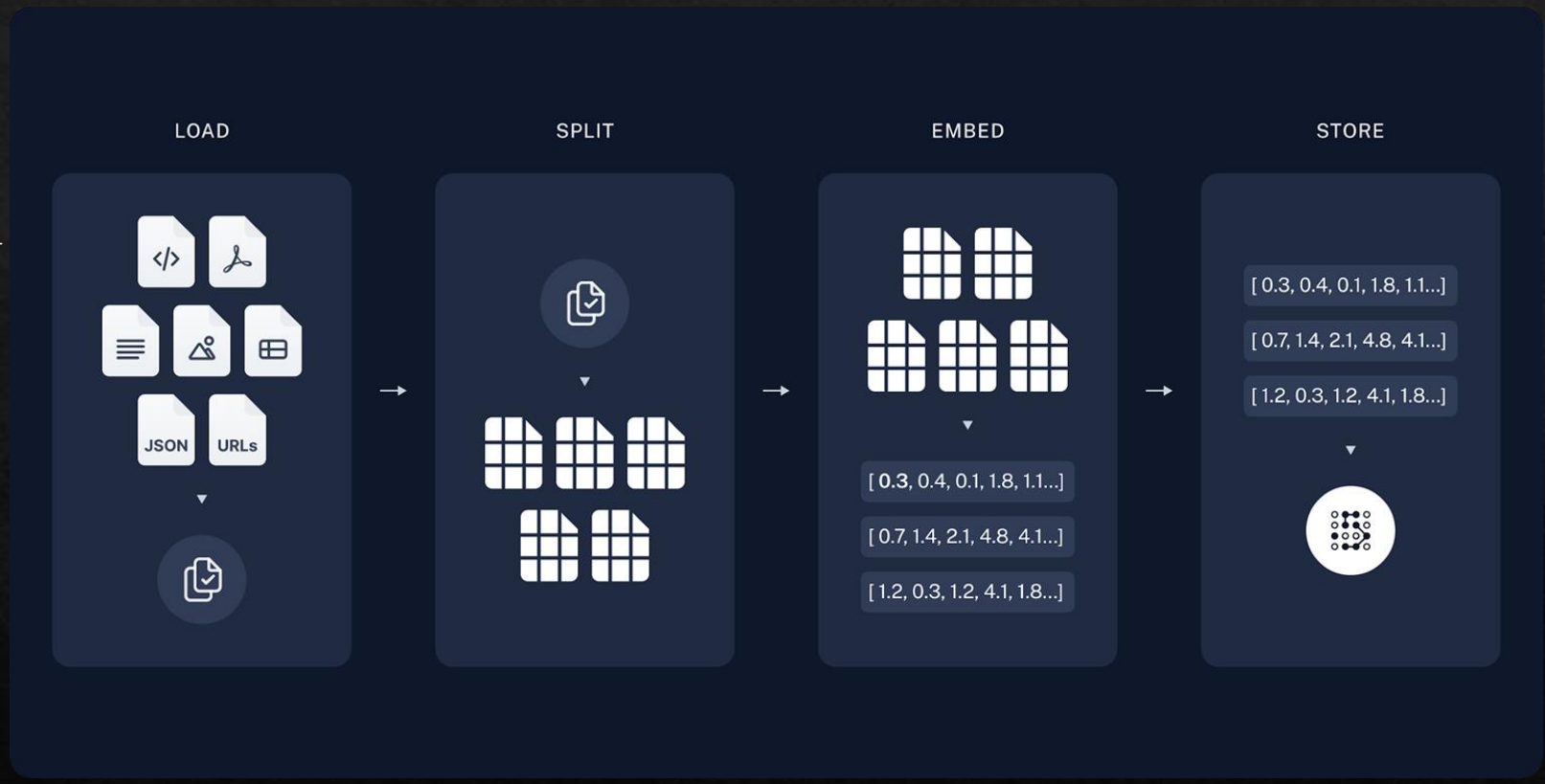
Jan Polanco Velasco

RAG

- Es una técnica para aumentar el conocimiento de un LLM
- Usa data adicional
- Una forma “alterna” al fine tuning
- Los LLMs tienen “conocimiento” sobre información pública
- Tienen fecha de entrenamiento
- Retrieval Augmented Generation

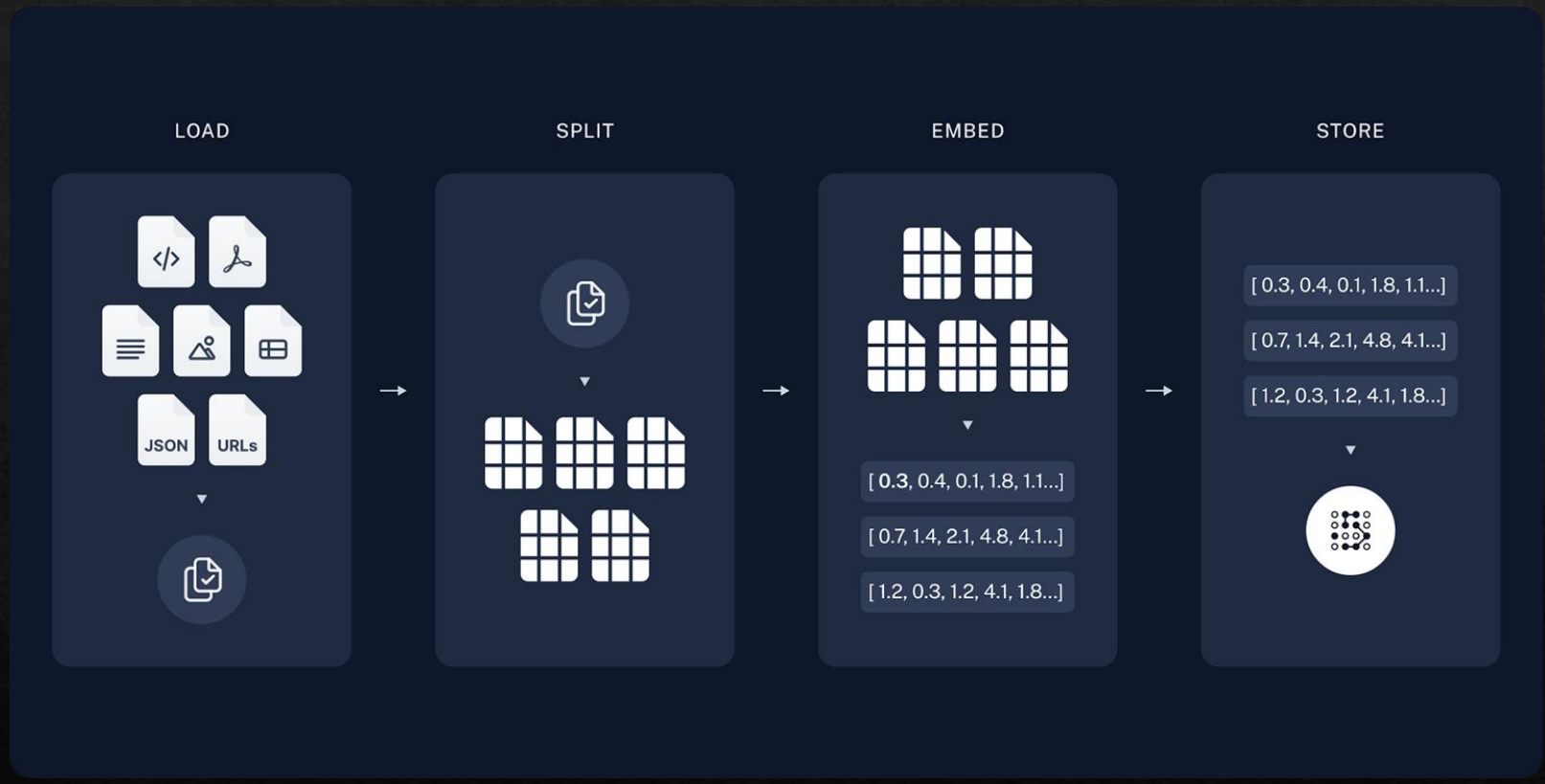
RAG

- Se usa RAG cuando se va a trabajar con data privada.
- Aumentar el “conocimiento del LLM” basado en ciertas especificaciones y necesidades.
- Un RAG tiene 2 bloques
- Indexación
- Retrieval and Generation



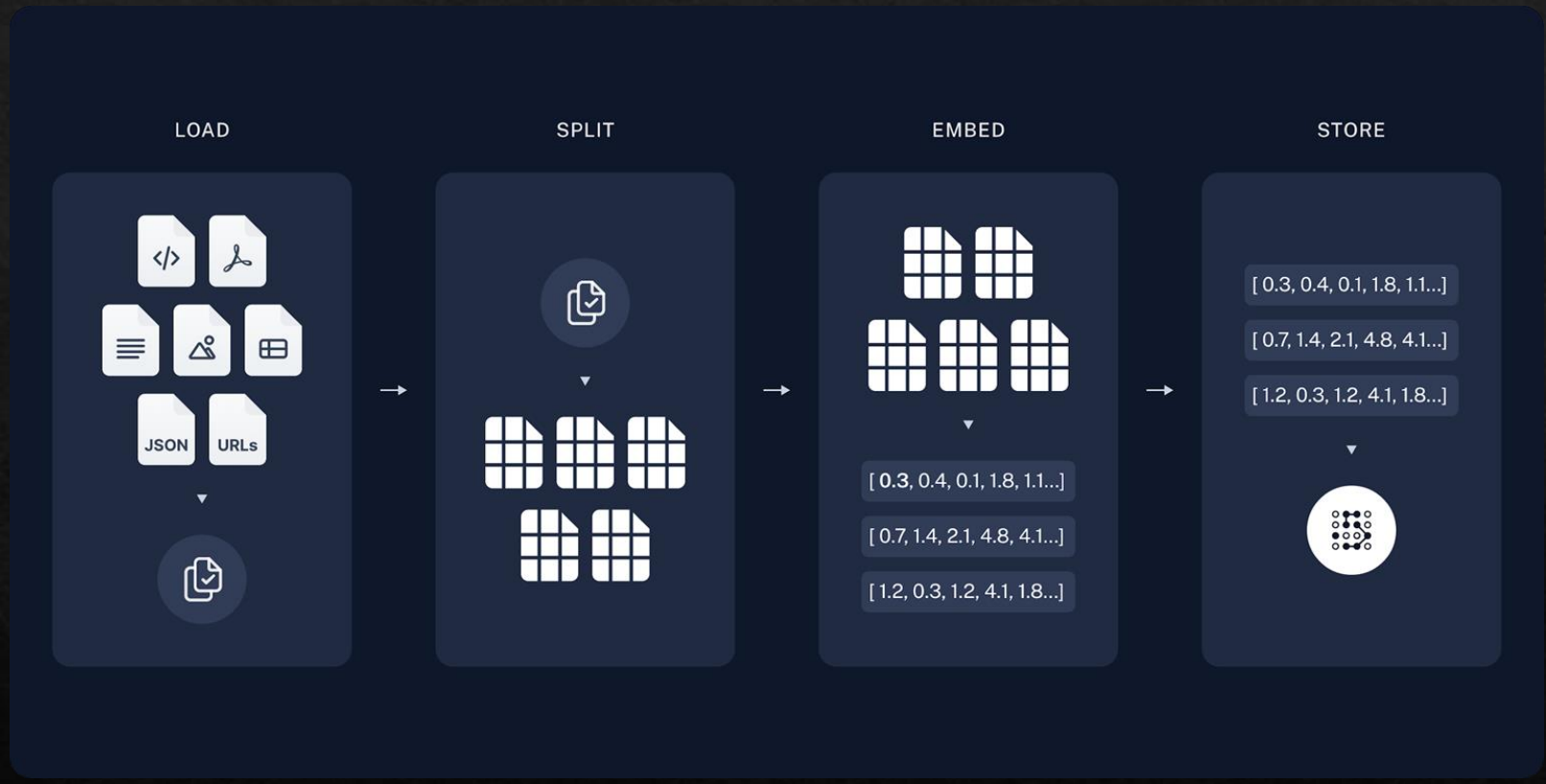
RAG

- Para la parte de indexación
- 1. Cargar nuestra data.
- 2. Split, acá se parte la data en chunks.
- 3. Almacenamiento, es necesario almacenar la informacion en algun lugar (nuestros chunks).
- NOTA: Entre mas grandes sean los chunks más difícil es la búsqueda.



RAG

- Para la parte de indexación
- 1. Cargar nuestra data.
- 2. Split, acá se parte la data en chunks.
- 3. Almacenamiento, es necesario almacenar la informacion en algun lugar (nuestros chunks).
- NOTA 2: El modelo se puede quedar sin contexto



RAG

- Para la parte de Retrieval and Generation
- 1. Dado un prompt de usuario splits relevantes son traídos desde el almacenamiento usando retriever