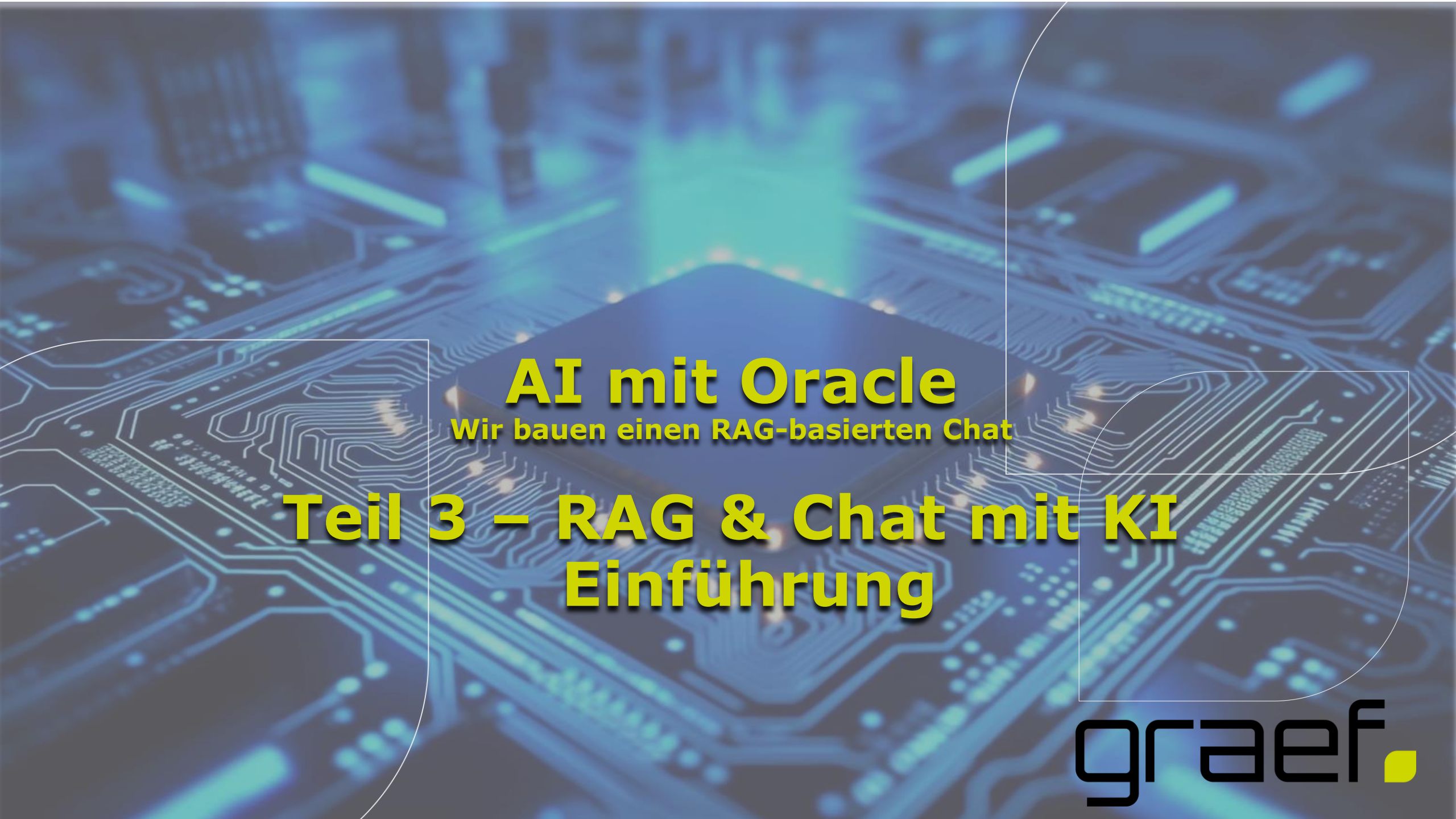


AI mit Oracle

Wir bauen einen RAG-basierten Chat

Teil 3 – RAG & Chat mit KI Einführung

graef.

Was bisher geschah...

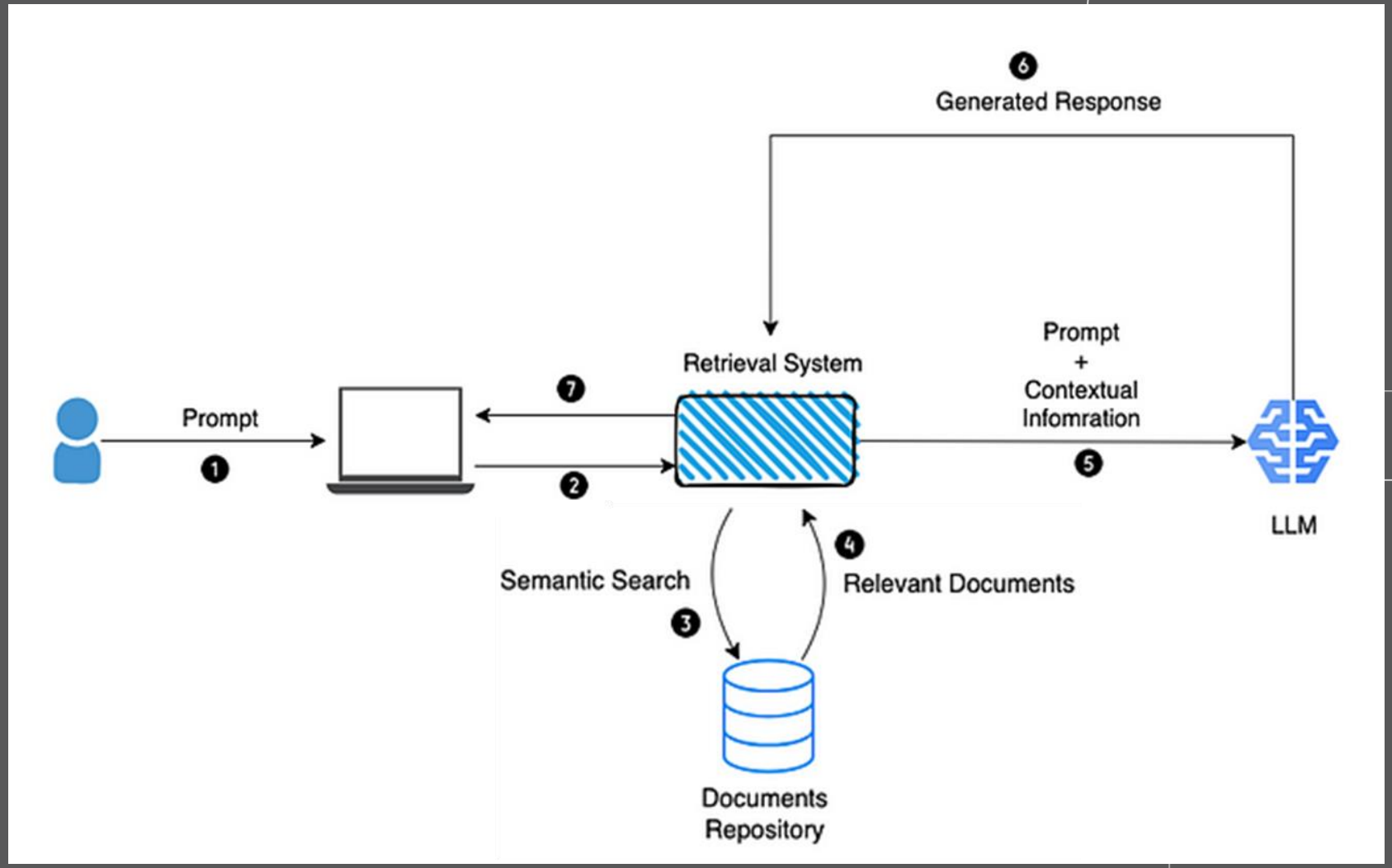
- **Chat auf Basis von ChatGPT für die Oracle-Datenbank gebaut**
- **Implementation rein PL/SQL + Nutzung OpenAI-API per UTL_HTTP**
- **Chat-Historie als Wissensbasis → iterative Kontextentwicklung im Chat möglich**

Was kommt jetzt?

- Integration von Spezialwissen mittels RAG = Retrieval Augmented Generation (Generieren [z.B. von Text] angereichert durch Abrufen [von Informationen]
- Konzept des RAG-Verfahrens

graef.

RAG – was ist das?



RAG Flow Diagram

Quelle: https://blog.gopenai.com/retrieval-augmented-generation-101-de05e5dc21ef

RAG – Kontextunterstützung des Chats

Konzeption

- Standard-RAG: Dokumenten-Teile als Wissensbasis
- Dokumentenbasis (Teil 1)
  - Datenstruktur Dokumente: Dokument←TextChunks
  - Chunking Verfahren
    - Via Oracle
    - Via Python
- Vektorsuche (Teil 2)
  - Grundsätzliche Vorüberlegungen
  - Verschiedene Verfahren zur Vektorisierung
    - Externer Service (z.B. ChatGPT in Oracle 23AI integriert)
    - Eigener REST-Service mit Python & Docker
    - ONNX-Integration in Oracle (23AI)
- Integration in den Chat (Teil 3)
  - Datenstruktur Chat: Chat←Historieneintrag←verwendete TextChunks

graef.

RAG – Kontextunterstützung des Chats

Viel Spaß bei den nächsten Teilen der Serie!

graef.

Graef Computer GmbH · Fallgatter 5 · 44369 Dortmund
Telefon: +49 231 222 429 - 99 · info@graef.com