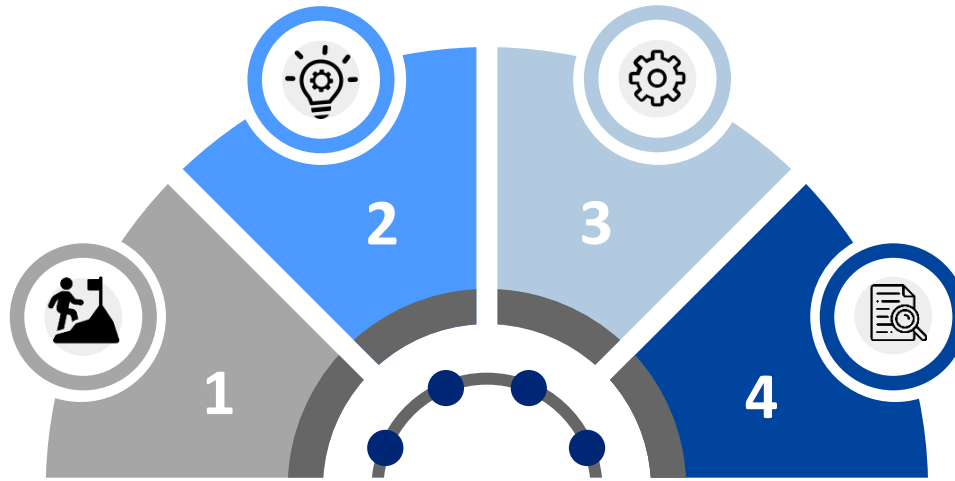# Case study #1 RAG case study

Solution: Implemented Retrieval Augmented generation through prompts enriched with the help of domain specific knowledge retrieval passed as context.

Parameters: GPT standards
Hardware: Cloud based AI optimized compute instance
Software: Python, Azure Open AI, Solr

Challenges: Achieving generative response with the help of domain specific knowledge base without needing to fine-tune LLM.

Overview: Created RAG search for retail specific response and NER

2

3

1

4

Start

Website

Query, Session ID

Search Query to Query log history Index

Search Index

Store Response with SessionId in Query log history Index

End

Construct prompt with Elastic Reaponse as context and Query

Construct Prompt with Query, Elastic Reaponse, and chat history

Azure Open AI

Gen AI Response for Context

Output renderred on separate panel