GenAI RAG Accelerator

Harness the power of Generative AI - Retrieval-Augmented Generation (RAG) for your business

What is RAG and Q for Business

Retrieval Augmented Generation (RAG) is a natural language processing technique that enables generative AI models to reference an authoritative knowledge base before generating an answer.

RAG can be used to build AI assistants tailored to a specific business domain. Amazon Q for Business enables the rapid creation, tuning and deployment of RAG solutions.

Self-Build RAG Challenges

Whilst tooling exists to enable organisations to build similar capabilities, self-build comes with challenges such as:

- Internal expertise in LLM architectures and frameworks needs to be developed
- Appropriate hardware and software needs to be purchased
- On-going maintenance and operational capabilities need to be in place
- Security and compliance requirements need to be met
- Data management, versioning and re-indexing need to be managed to ensure ongoing accuracy of the assistant.

GenAI RAG Accelerator

Enables customers to rapidly pilot and explore the benefits of Generative AI for their business by leveraging off the shelf templates and our expertise in building custom AI solutions.

Post pilot we support secure and compliant rollout of the RAG solution across the organisation.

Who is it for?

Organisations with internal domain specific knowledge bases; for example, standard operating procedures and market intelligence. Example verticals include:

- Financial services such as Insurance, Banking, Investments, Capital Markets
- Life Sciences and manufacturing
- Energy

What does it cost?

By leveraging AWS Funding programs, we minimize or eliminate all customer costs for the Accelerator.

Key Benefits of RAG with Q for Business

Boost team productivity - Q for Business streamlines workflow by summarising documents, generating drafts, conducting research, or running comparative analysis.

Tailored to your needs - Q connects to company information, allowing it to generate content and take actions that are relevant to your business.

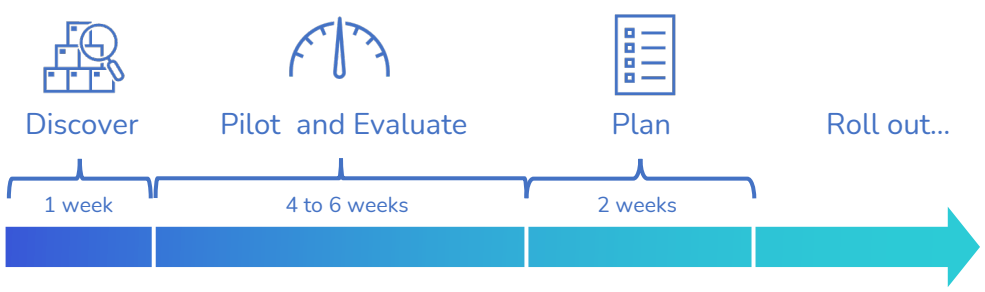
Secure and compliant - Q understands and respects existing role and permission structures and boundaries, enabling customers to meet their most stringent enterprise compliance needs.

Dramatically Reduce code and infrastructure - Compared to self-build, a Q implementation has a far lower implementation and maintenance overhead, reducing TCO.

Always Improving - The LLM technology backing Q is constantly evolving, meaning that improvements can be incrementally deployed at low cost.

Sustainability - By using centrally trained and managed models, as opposed to self-training Q improves you companies carbon footprint.

How it works



The RAG accelerator beings with an initial discovery workshop which typically takes one day. The aim of the workshop is to identify high business value use cases. Following the workshop, we determine the which of the identified use cases to pilot based on the availability of data, security and compliance requirements and end user needs. We also determine a set of objective business KPIs to measure during the pilot. In total this phase takes takes approximately one week.

During the pilot phase, we configure and deploy the Q RAG application(s) along with the associated infrastructure required to securely connect to data sources and end users and to measure the identified KPIs. Deployment of the pilot typically takes two to three weeks. We then allow time for users to engage with the assistant to collect feedback.

Objective KPIs are measured and, along with subjective user feedback we evaluate the business impact of the pilot. Based on these measurements we determine a rollout plan to move to a full production implementation. The plan typically addresses issues such as data security, classification, versioning and role-based access.

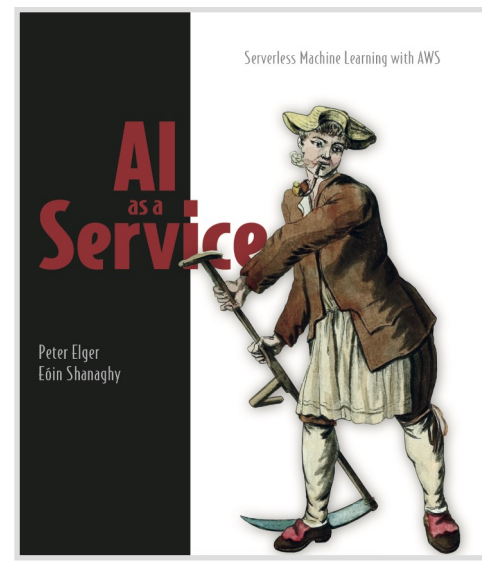
Why fourTheorem?

We wrote the book on it!

AI as a Service is a detailed guide to operationalizing Cloud Native AI Services on AWS

As an advanced consulting partner, we have architected and delivered systems for our customers that harness the power of AWS AI services.

“A practical approach to real-life AI smartly based on a serverless approach. Enlightening!”



AUGMENTED UNDERWRITING

Our Augmented Underwriting RAG assistant has been trained on 24 years of Lloyds of London public market bulletins

Try it out: https://augmentedunderwriting.com/

aws PARTNER Advanced Tier Services

- Advanced Tier Services
- AWS Lambda Delivery
- AWS ECS Delivery
- Migration Services
- Well Architected

Have a project you’d like to discuss or want to know more? +353 21 206 3101 hello@fourtheorem.com www.fourtheorem.com