BigCheese

aws PARTNER Premier Tier Services

# Start Now GenAI Agent Pack SOW

## AI/ML Data

*Version 1.0*

*May 9, 2024*

*By: Uriel Agustín Grosso - Pre-Sales Solutions Architect*

# Index

# Introduction

**Make every conversation count**

At BigCheese, an AWS Premier Partner, we believe the true value of generative AI lies not only in its ability to answer, but in its ability to **understand, adapt, and unlock** real value at every user touchpoint.

Today, the most agile organizations don't just automate tasks, they **talk to their systems,** learn from interactions, and continuously improve both customer and employee experience.

That’s why we apply our **Think Big, Start Small** approach to the world of conversational GenAI:

- **Think Big** is about imagining an organization where customers, employees, and systems connect naturally through intelligent assistants, available 24/7, capable of answering complex questions, performing tasks, and learning from every interaction.

- **Start Small** means launching with a customized GenAI Agent, integrated with your internal systems, connected to your knowledge sources, and deployed on a secure, scalable, and efficient AWS architecture.

With the **Start Now GenAI Agent** package, we offer to design, train, and deploy a next-generation conversational assistant in just 8 weeks. A focused project, with measurable results, and a design built to scale.

This approach brings three key advantages:

- **Answers that matter**: The Agent responds based on your own knowledge (RAG), with context, precision, and traceability.
- **Real integration**: The assistant connects with your systems, documents, and workflows—going beyond generic responses.
- **Adoption strategy**: We design a specific use case (customer-facing, internal, or support) so you can test quickly and scale based on real results.

GenAI transformation doesn't start in a lab or with isolated pilots.

**Start where the real value is: in the actual conversations of your business.**

## Why Build Your GenAI Agent with BigCheese?

At BigCheese, we're an AWS Premier Partner and one of the few certified partners in Latin America with AWS's official **Generative AI Competency**, positioning us as a leader in conversational AI solutions across the region.

Our team isn't just certified in GenAI, we also bring expertise in:

- **Machine Learning Specialty**
- **AI/ML Practitioner**
- **Data Engineering on AWS**
- **Data Analytics Specialty**

And most importantly: we have a **dedicated GenAI Squad**, exclusively focused on designing, training, and deploying custom conversational assistants. These assistants are integrated with internal systems and aligned with the real workflows of your business..

Building your GenAI Agent with BigCheese means working with a team that understands it's not just about "talking", it's about responding with **context**, connecting with **your data**, and creating **experiences that learn and scale**.

We design with a focus on **security, efficiency, and adoption**. We don't deliver disconnected proof-of-concepts, we deliver conversational solutions ready to integrate, evolve, and generate real impact.

## Who should use this service?

The **Start Now GenAI Agent** is designed for organizations looking to improve **customer or employee experience** through secure, scalable, and tangible conversational AI solutions.

It's ideal for companies that:

- **Handle a high volume of repetitive queries**

  Support teams that receive hundreds of daily questions, many of them simple, documented, or repetitive, and want to automate intelligently.

- **Want faster and higher-quality responses**

Organizations dealing with slow or inconsistent responses that want to offer a 24/7, contextual, and professional experience without depending solely on human agents.

- **Want to get started with GenAI but don't know where**

  Businesses eager to explore generative AI but need a focused, low-risk, high-impact use case to validate its potential.

- **Need to connect a bot to their own data**

  Companies with internal documentation, manuals, policies, or FAQs who want the Agent to respond based on their own content using **RAG (Retrieval-Augmented Generation)** techniques.

- **Are scaling or undergoing digital transformation**

  Growing organizations or those in the process of digitalization that see GenAI not only as a Agent, but as a tool to **boost operational efficiency and unlock new services.**

## Objective

The objective of the **Start Now GenAI Agent** is to help organizations implement an intelligent conversational assistant that responds accurately based on their **own internal knowledge**, improving user experience and reducing operational load.

In just 8 weeks, we design and deploy a GenAI-based solution with a **RAG (Retrieval-Augmented Generation)** approach that enables:

- Answering real questions using internal documents, policies, manuals, or databases

- Automating repetitive queries in internal processes or customer service
- Integrating the Agent with key services such as Amazon Bedrock, S3, Redshift, or proprietary systems
- Laying the groundwork to scale across multiple channels (WhatsApp, Teams, web, etc.) and other AI use cases

**Turn your organization's knowledge into a competitive advantage, available 24/7.**

## What do we need for a successful GenAI Agent?

To ensure that the **Start Now GenAI Agent** delivers real impact within 8 weeks, we require:

- **Access to relevant internal knowledge**

  Updated documentation (manuals, PDFs, policies, instructions, FAQs, etc.) that represents the knowledge you want the Agent to access. This can also include databases or internal systems accessible via API or queries.

- **Available functional and technical stakeholders**

  People who understand the current processes, documents, or systems and can help validate responses, refine the approach, and define what constitutes a "correct" or "useful" answer.

- **Clear and prioritized use cases**

Clarity on what kind of queries we want to automate: customer-facing, internal support, onboarding, product questions? Having a focused goal allows us to build a real and measurable conversational experience from day one.

- **Willingness to test and provide feedback**

  We need to validate real Agent responses, fine-tune tone, scope, accuracy, and integrations. Demo and collaborative feedback sessions are key to ensuring a useful end result.

- **Commitment to Adoption**

  The real value lies not just in deploying the assistant, but in **actually using it.** That's why we work with companies ready to integrate it into their channels, promote it to users, and build upon that first use case.

## Service Scope

The scope includes:

**Initial assessment and business understanding**

- Identify processes or areas with high volumes of repetitive queries
- Review knowledge sources: documents, internal databases, APIs, FAQs
- Define the prioritized use case (customer service, internal support, onboarding, etc.)

- Map objectives, frequent questions, and language/response constraints

**AWS Architecture design for GenAI + RAG**

- Architecture based on Amazon Bedrock + Amazon S3, RDS, or OpenSearch as data sources
- Define document extraction, processing, and indexing workflows
- Design RAG pipeline (vectorization, embeddings, query, and response generation)
- Define security layers, access control, and traceability

**Technical implementation**

- Integrate with Amazon Bedrock and embedding models (Titan, Cohere, Claude, etc.)
- Index documents and build context-retrieval flows
- Design prompts, tone restrictions, and fallback logic
- Validate response accuracy and control Agent output
- Configure access channel: web, internal tool, API, Teams, etc. (based on use case)

**Documentation and adoption preparation**

- Functional manual: how to upload new documents, interpret responses, and fine-tune the Agent

- Technical documentation: architecture, indexing, flows, and applied adjustments
- Final demo of the Agent operating with real client data
- Next steps guide: continuous training, new channel connections, expansion strategy

**Knowledge Transfer**

- Final functional and technical workshop
- Deliverable documentation package
- Q&A session and best practices for post-launch evolution

## Success criteria

To consider the **Start Now GenAI Agent** service successful, the following technical and functional criteria must be met and validated with the client:

- **Functional GenAI Agent using internal knowledge**

  The assistant must be deployed and capable of answering real questions using the client's internal documents, FAQs, or knowledge bases, with **at least 80% validated response accuracy**.

- **Operational RAG Architecture**

  The complete retrieval-augmented generation pipeline must be working: document upload, vectorization, embedding storage, and integration with Amazon Bedrock or another compatible LLM.

- **Custom interaction flow, fully tested**

  The Agent must demonstrate conversational behavior aligned with the defined use case (tone, limits, context, fallback), and handle multiple questions within the same thread.

- **Knowledge transfer workshop delivered**

  The client team must participate in a final session explaining how to add new information, fine-tune the assistant's behavior, and scale the solution further.

- **Technical documentation delivered**

  A documentation package must be delivered, including the deployed architecture, service configurations (Bedrock, S3, embeddings), prompts, flows, and future evolution recommendations.

- **Validated usage channel**

  The Agent must be fully functional on the selected channel (web, Teams, internal API, etc.), with **controlled access** and **usage traceability**.

## Deliverables

Throughout the 8-week implementation, the client will receive a set of deliverables that ensure the Agent based on GenAI and deployed on AWS is fully operational, functional, and well-documented:

- **Deployed and Functional GenAI Agent**

  Conversational assistant ready to use, based on Amazon Bedrock or an equivalent service, integrated with at least one internal source (documents, database, or FAQ), and accessible through the defined channel (web, Teams, API, etc.).

- **Indexed and prepared document set**

  Ingestion and processing of agreed-upon files/documents, with a fully functional embedding pipeline for contextual queries using RAG (Retrieval-Augmented Generation) architecture.

- **Full retrieval + generation flow**

  Operational pipeline that performs semantic retrieval, applies the prompt, and generates answers using an LLM model. Includes fallback configuration, tone control, and query logic.

- **Validated Functional Testing**

  Conversational scenarios defined and validated with the client's functional users, including at least 10 representative queries aligned with the prioritized use case.

- **Complete Technical Documentation**

A detailed manual including:

  - Deployed architecture.

- Service configuration (Bedrock, S3, Redshift, API Gateway, etc.).
- Prompting setup.
- Embedding and retrieval workflows.
- Recommendations for future adjustments.

- **Knowledge transfer workshop**

  Final session with technical and functional users to explain usage, maintenance, and Agent evolution. Focus on scalability, adding new documents, and customization.

## Involved Team

To ensure the success of this initiative, BigCheese will assign a technical team with proven expertise in generative AI, natural language processing, and conversational solutions on AWS, composed of the following roles:

- **Project Manager**

  Responsible for planning, monitoring, and coordinating with the client's teams. Ensures on-time delivery, manages dependencies, and leads progress reviews and functional validations.

- **Solutions Architect (GenAI Specialist)**

  Designs the Agent's architecture, defines AWS components (such as Bedrock, S3, Lambda, API Gateway), ensures best practices for

security, scalability, and governance. Acts as the bridge between business goals and the technical solution.

- **AI/ML Engineer**

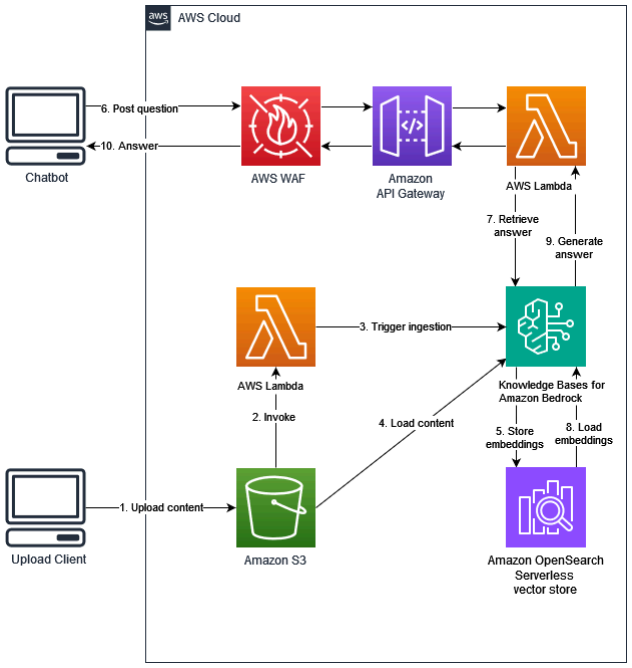
  In charge of configuring the base model in Amazon Bedrock, building and refining prompts, integrating the RAG logic (retrieval augmented with embeddings), validating response accuracy, and fine-tuning the Agent's conversational behavior.

- **Data Engineer**

  Processes and indexes internal sources (PDFs, FAQs, manuals, SQL databases), builds embedding pipelines, and ensures the Agent has structured, efficient access to internal knowledge.

This team works collaboratively with the client's technical and functional stakeholders to ensure the Agent delivers accurate, secure, and user-friendly responses from day one.

# Reference architecture




# Action Plan

| Phase | Activity | Estimated Hours |
|---|---|---|
| **Kickoff & Discovery** | Kickoff meeting, use case definition, document identification, channels and objectives review | 8 |
| | Review of provided documentation, categorization and prioritization | 6 |
| **GenAI + RAG Architecture Design** | Architecture design (Bedrock, S3, embeddings, API) | 4 |

| | | |
|---|---|---|
| | Definition of RAG flows, initial prompts, and response validation scheme | 4 |
| **Knowledge processing & loading** | Cleaning, adaptation, and conversion of documents for use | 12 |
| | Indexing sources into vector store + embeddings | 14 |
| **Technical implementation** | Bedrock configuration, RAG orchestration and response generation | 30 |
| | Integration with defined channels (web, Teams, API, etc.) | 12 |
| **Functional validation & tuning** | Response testing, user validation, prompt refinement and fallback logic | 12 |
| | Precision, tone, and context adjustments | 6 |
| **Documentation & knowledge transfer** | Technical manual covering architecture, prompts, usage and maintenance | 8 |
| | Final workshop with client's technical and functional teams | 4 |
| **Project management & coordination** | Progress tracking, planning, control meetings, client communication | 20 |

**Total Assigned Hours: 140 hours**

## Cronograma estimado

| | Month 1 | | | | Month 2 | | | |
|---|---|---|---|---|---|---|---|---|
| **Phase** | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** | **S8** |
| Kickoff & Discovery | | | | | | | | |
| GenAI + RAG Architecture Design | | | | | | | | |
| Knowledge Processing & Loading | | | | | | | | |
| Agent Technical Implementation | | | | | | | | |
| Functional Validation & Tuning | | | | | | | | |
| Documentation & Handover | | | | | | | | |
| Project Management & Coordination | | | | | | | | |

## Budget

Total team hours: **140 hours**

**Total cost:** USD 11,200 + VAT

**Special offer via AWS Marketplace: 50% discount:**

**USD 5,600 + VAT**

## What comes after the PoC?

The **Start Now GenAI Agent** is just the beginning. This proof of concept is designed not only to deliver real value through the first use case, but also to lay the foundation for a scalable, integrable, and evolving solution.

Once the project is complete, organizations can:

- **Scale the assistant to new channels**

  Integrate the Agent with high-impact platforms like WhatsApp, Teams, Slack, internal apps, websites, or mobile apps to extend its reach.

- **Expand the knowledge base**

  Add new documents, systems, and dynamic sources (SQL databases, APIs, CRMs) to cover more topics, with deeper knowledge and automatic updates.

- **Connect with internal processes**

  Link the assistant to ticketing systems, HR tools, tech support, customer databases or approval workflows so it can **inform and take action.**

- **Measure and continuously improve**

  Apply usage, accuracy, satisfaction, and adoption metrics to train new versions, refine prompts, and evolve its intelligence based on evidence.

- **Standardize it as an internal or commercial copilot**

  Evolve the Agent into a **corporate copilot** to assist in daily tasks, onboarding, sales support, document management, and beyond.

At **BigCheese**, we support you every step of the way, with a team specialized in GenAI, AWS architecture, and business-centered solutions.

**The PoC proves it works. What comes next proves it transforms.**