Challenge AWS AI Practitioner Certified

AWS COMMUNITY CARTAGENA

BIENVENIDOS

Semana 4: Optimización de FM, seguridad y gobierno

AWS COMMUNITY
CARTAGENA

Acerca de mí

Esposo y padre de 2 hijos, Ángel y Sara
Ingeniero de sistemas, especialista en telecomunicaciones, MBA
Trabajo en el campo de la tecnología hace
más de 20 años.
Áreas de especialización

- Arquitectura TI
- DevOps
- Site Reliability Engineering
- Arquitectura multi cloud (AWS, Azure, OCI)
- https://www.linkedin.com/in/mario-rodrigo-serrano-pineda/
- E-mail: marosepi2020@gmail.com
- https://medium.com/@marioserranopineda

AGENDA

CloudOps Guild
Together, towards mastery in Cloud and DevOps

1. Optimización de modelos fundacionales
2. Seguridad, cumplimiento y gobierno de soluciones IA

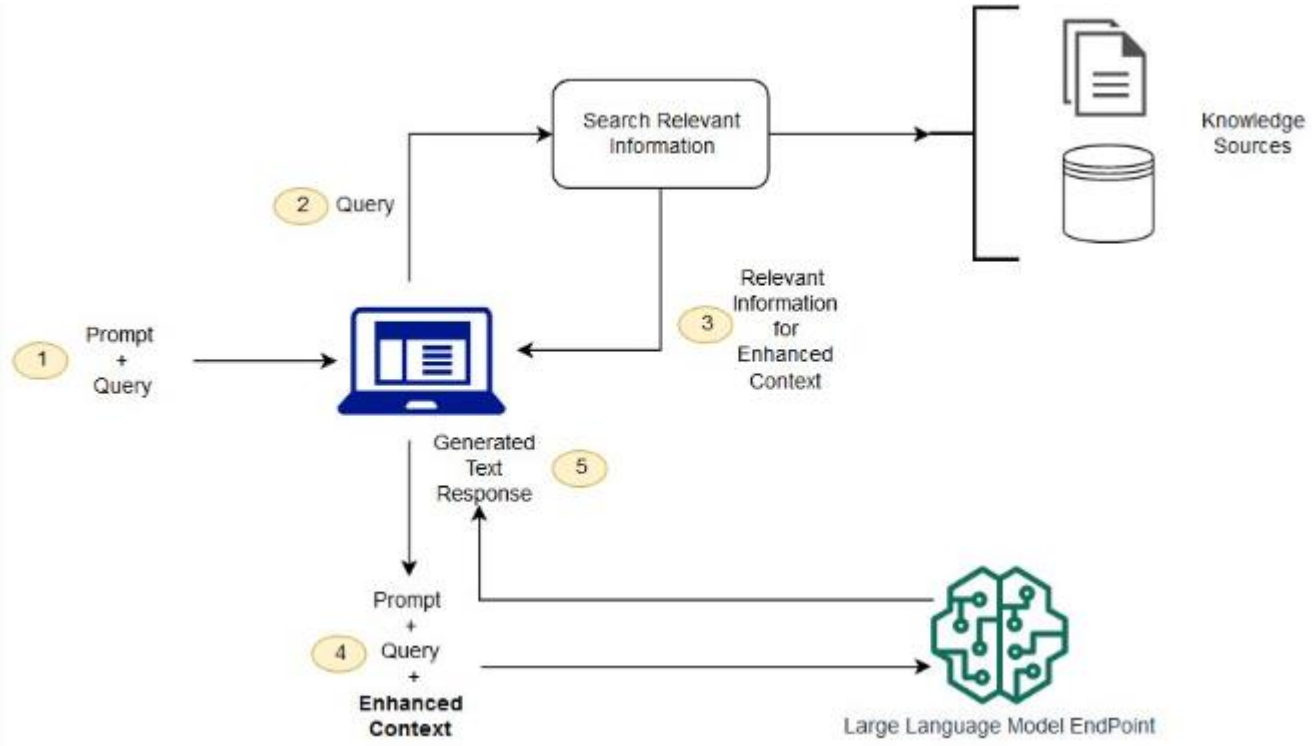
AWS COMMUNITY
CARTAGENA

AWS COMMUNITY
CARTAGENA

1. Optimización de modelos fundacionales

¿Qué es la generación aumentada por recuperación?

Retrieval-Augmented Generation (RAG) es el proceso de optimización de la salida de un modelo lingüístico de gran tamaño, de modo que haga referencia a una base de conocimientos autorizada fuera de los orígenes de datos de entrenamiento antes de generar una respuesta. Se trata de un método rentable para mejorar los resultados de los LLM de modo que sigan siendo relevantes, precisos y útiles en diversos contextos.



CloudOps Guild
Together, towards mastery in Cloud and DevOps

AWS COMMUNITY CARTAGENA

Caso empresarial

Centro de soporte de la empresa UnaEmpresa S.A.

UnaEmpresa es un proveedor de telecomunicaciones que ofrece servicios de telefonía e internet.
Por lo general, los clientes que tienen problemas se contactan con el centro de soporte de UnaEmpresa por teléfono, un proceso que es costoso e ineficiente.
Para agilizar el servicio de soporte, UnaEmpresa está adoptando soluciones en línea. En un principio, creó una sección de preguntas frecuentes en su sitio web para resolver los problemas más comunes. A pesar de esto, el volumen de tickets de soporte en línea sigue siendo alto. Para reducir aún más la carga de trabajo del personal y mejorar la eficiencia del servicio al cliente, UnaEmpresa está explorando el potencial de la inteligencia artificial (IA) generativa para desarrollar un chatbot. El chatbot sería capaz de guiar a los clientes, responder a los problemas más comunes y realizar ciertas tareas de forma autónoma, como pedir un teléfono nuevo, actualizar el plan telefónico para obtener más datos 5G, etc.
UnaEmpresa tiene el objetivo de reducir la cantidad de tickets en línea en un 70 % al implementar el chatbot. La empresa también supervisará el nivel de satisfacción de los clientes a través de una encuesta que les proporcionará luego de resolver los problemas. Con la incorporación del chatbot, la empresa desea alcanzar un nivel de satisfacción de, como mínimo, 4 sobre 5.

Diagrama de arquitectura del sistema de soporte de UnaEmpresa.

Caso empresarial

Centro de soporte de la empresa UnaEmpresa S.A.

Para crear el chatbot, UnaEmpresa debe elegir un modelo fundacional (FM). El modelo necesitará habilidades de procesamiento de lenguaje natural (Natural Language Processing, NLP), además de una integración con el backend de UnaEmpresa para la automatización.
UnaEmpresa debe elegir un modelo de lenguaje grande (Large Language Model, LLM) porque estos FM tienen la capacidad de entender y procesar el lenguaje natural. Los LLM se entrenan con una gran cantidad de datos públicos. Esto es excelente para lograr la comprensión general del lenguaje. Sin embargo, no es la mejor opción para responder a las solicitudes específicas que realizan los clientes sobre los servicios de UnaEmpresa. Por lo tanto, UnaEmpresa necesita encontrar una forma de incorporar datos y una base de conocimientos propios para darle más contexto al chatbot y mejorar la exactitud de las respuestas.
Estos datos complementarios pueden provenir de registros de chat, de tickets de soporte gestionados anteriormente o, incluso, de grabaciones de llamadas de soporte. Hay que recopilar, anonimizar y limpiar los datos para poder incorporarlos a una base de conocimientos que, a su vez, pueda ser utilizada por el chatbot.
Por último, dado que UnaEmpresa está dispuesta a que el chatbot gestione algunas tareas de forma autónoma, este debe poder iniciar funciones adicionales capaces de modificar los parámetros de las cuentas de los clientes.

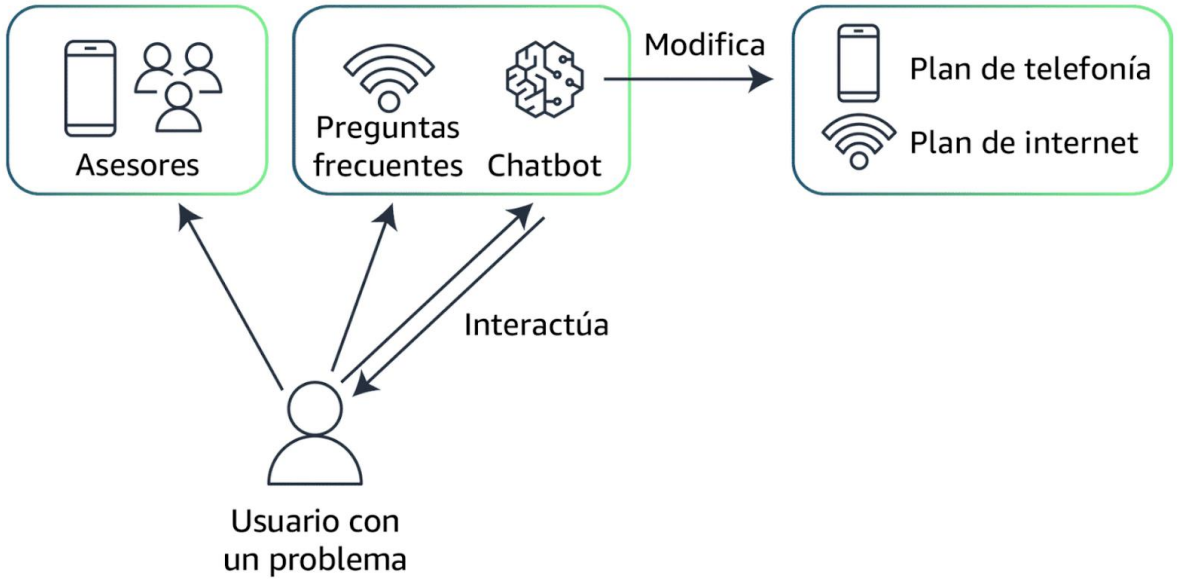


Diagrama de arquitectura del sistema de soporte de UnaEmpresa.

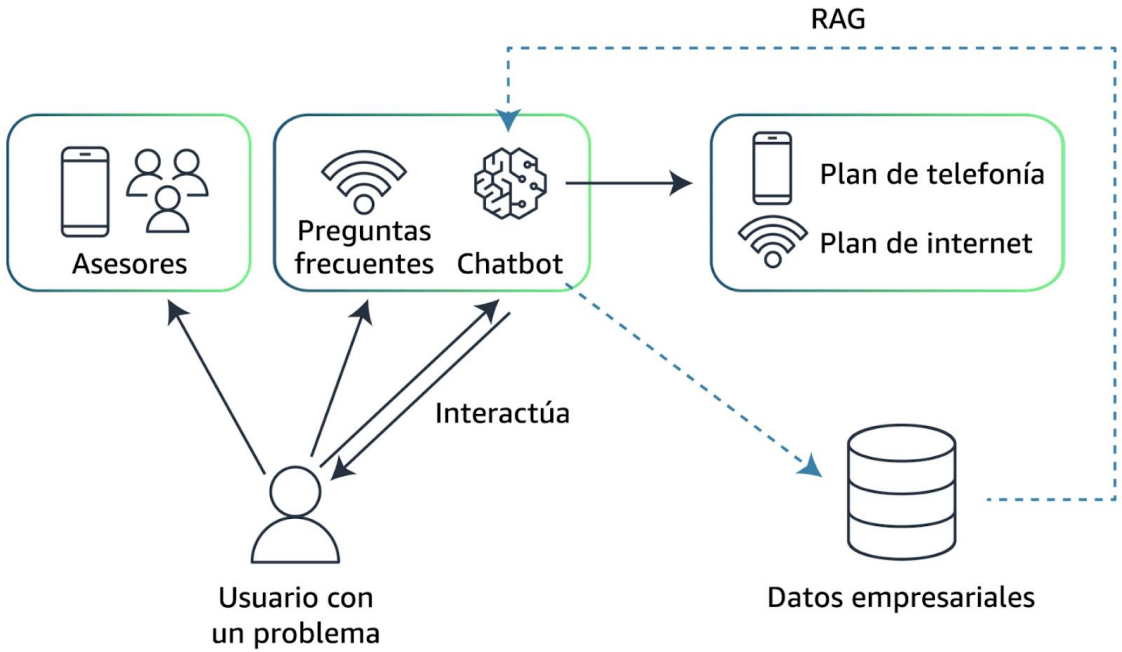
CloudOps Guild
Together, towards mastery in Cloud and DevOps

AWS COMMUNITY CARTAGENA

Caso empresarial

Centro de soporte de la empresa UnaEmpresa S.A.

Para crear el chatbot, UnaEmpresa debe elegir un modelo fundacional (FM). El modelo necesitará habilidades de procesamiento de lenguaje natural (Natural Language Processing, NLP), además de una integración con el backend de UnaEmpresa para la automatización.
UnaEmpresa debe elegir un modelo de lenguaje grande (Large Language Model, LLM) porque estos FM tienen la capacidad de entender y procesar el lenguaje natural. Los LLM se entrenan con una gran cantidad de datos públicos. Esto es excelente para lograr la comprensión general del lenguaje. Sin embargo, no es la mejor opción para responder a las solicitudes específicas que realizan los clientes sobre los servicios de UnaEmpresa. Por lo tanto, UnaEmpresa necesita encontrar una forma de incorporar datos y una base de conocimientos propios para darle más contexto al chatbot y mejorar la exactitud de las respuestas.
Estos datos complementarios pueden provenir de registros de chat, de tickets de soporte gestionados anteriormente o, incluso, de grabaciones de llamadas de soporte. Hay que recopilar, anonimizar y limpiar los datos para poder incorporarlos a una base de conocimientos que, a su vez, pueda ser utilizada por el chatbot.
Por último, dado que UnaEmpresa está dispuesta a que el chatbot gestione algunas tareas de forma autónoma, este debe poder iniciar funciones adicionales capaces de modificar los parámetros de las cuentas de los clientes.



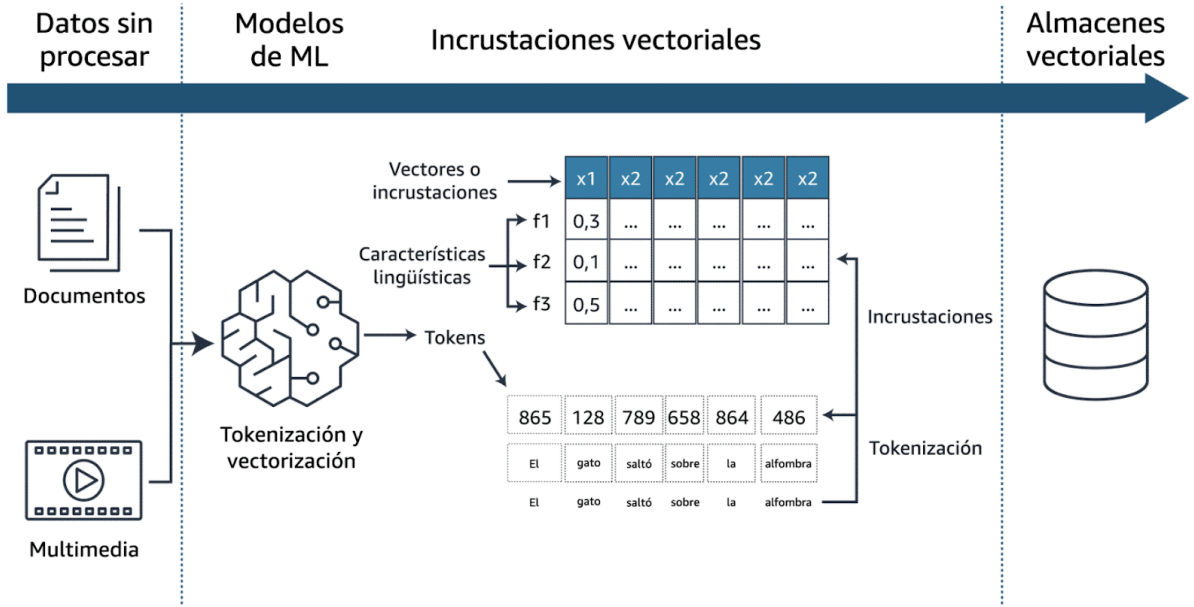
CloudOps Guild
Together, towards mastery in Cloud and DevOps

AWS COMMUNITY CARTAGENA

Del conjunto de datos a las incrustaciones vectoriales (Vectors Embeddings)

CloudOps Guild
Together, towards mastery in Cloud and DevOps

Las empresas recopilan grandes cantidades de datos internos como documentos, presentaciones, manuales de usuario, informes y resúmenes de transacciones, todos ellos desconocidos para el FM. Cuando estos modelos ingieren y utilizan orígenes de datos empresariales, adquieren conocimientos específicos del área, lo que les permite producir respuestas personalizadas y sumamente relevantes que satisfacen las necesidades empresariales.



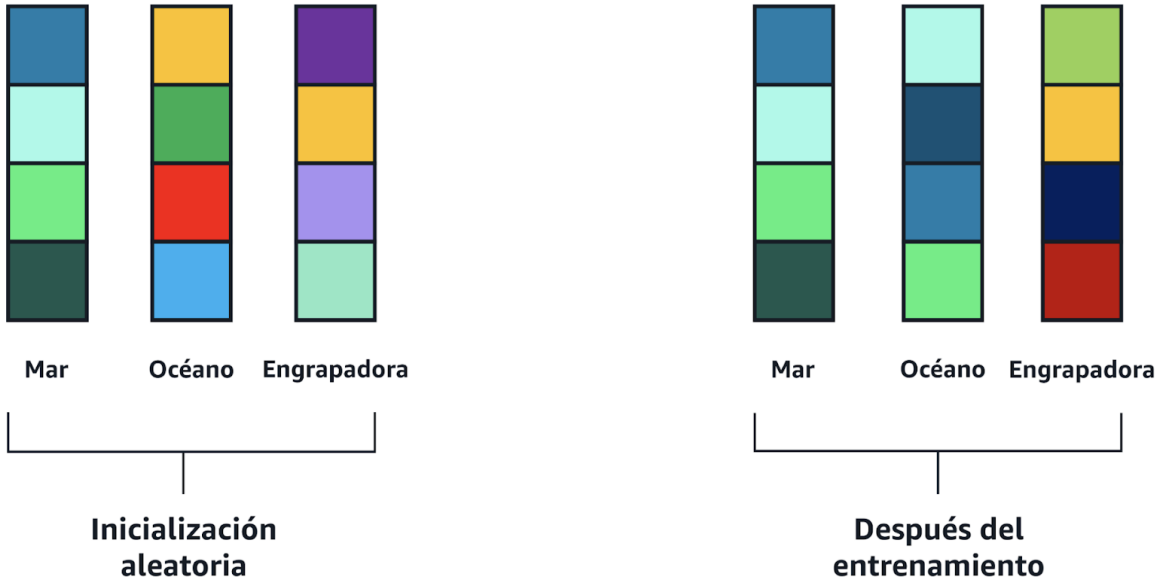
Los conjuntos de datos empresariales, como documentos, imágenes y audio, se pasan a los modelos de ML como tokens y se vectorizan.

AWS COMMUNITY
CARTAGENA

Del conjunto de datos a las incrustaciones vectoriales (Vectors Embeddings)

CloudOps Guild
Together, towards mastery in Cloud and DevOps

Dos palabras relacionadas tendrán incrustaciones similares.
Un ejemplo de estas dos palabras son mar y océano.



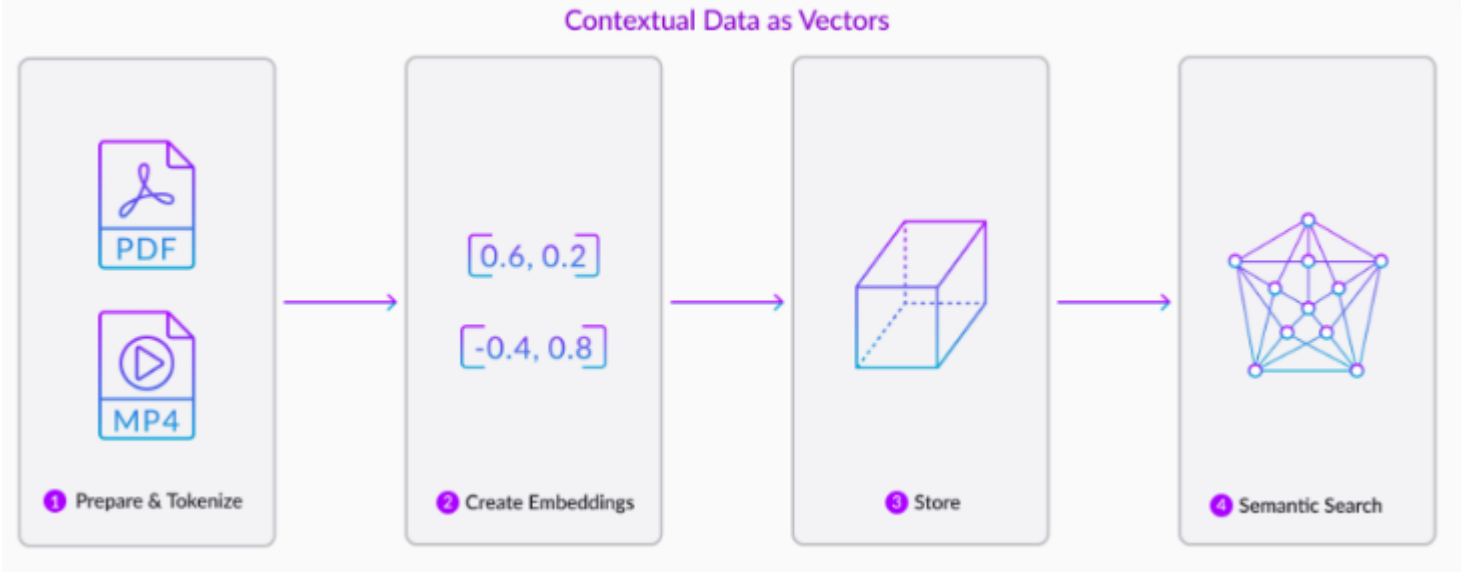
Las palabras que se relacionan entre sí tendrán incrustaciones similares.

AWS COMMUNITY
CARTAGENA

Bases de datos vectoriales en AWS

CloudOps Guild
Together, towards mastery in Cloud and DevOps

- Amazon OpenSearch Service (aprovisionada)
- Amazon OpenSearch sin servidor
- Extensión pgvector en Amazon Relational Database Service (Amazon RDS) para PostgreSQL
- Extensión pgvector en Amazon Aurora, edición compatible con PostgreSQL
- Amazon Kendra



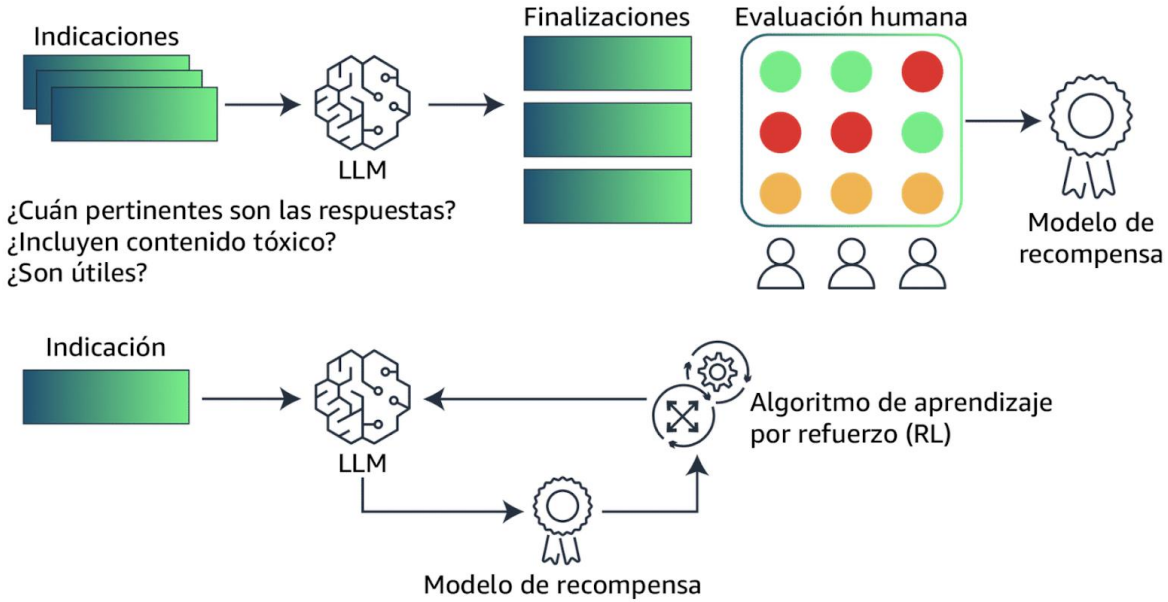
AWS COMMUNITY
CARTAGENA

Ajuste preciso

Los diferentes enfoques de ajuste preciso

•**Ajuste de instrucciones:** este enfoque implica volver a entrenar el modelo con un conjunto de datos nuevo que incluye indicaciones seguidas de los resultados deseados.

•**Aprendizaje por refuerzo a partir de la retroalimentación humana (Reinforcement Learning from Human Feedback, RLHF):** este enfoque es una técnica de ajuste preciso en la que el modelo inicialmente se entrena mediante el aprendizaje supervisado para predecir respuestas similares a las humanas.



CloudOps Guild
Together, towards mastery in Cloud and DevOps

AWS COMMUNITY CARTAGENA

AWS COMMUNITY
CARTAGENA

2. Seguridad, cumplimiento y gobernanza

Orientación estratégica para la seguridad, la gobernanza y el cumplimiento

Seguridad: Garantizar que se mantengan la confidencialidad, la integridad y la disponibilidad de los datos y los activos de información y la infraestructura de la organización.
Esta función suele denominarse seguridad de la información o ciberseguridad en una organización.

Gobernanza: Garantizar que una organización pueda agregar valor y gestionar el riesgo en la operación del negocio.

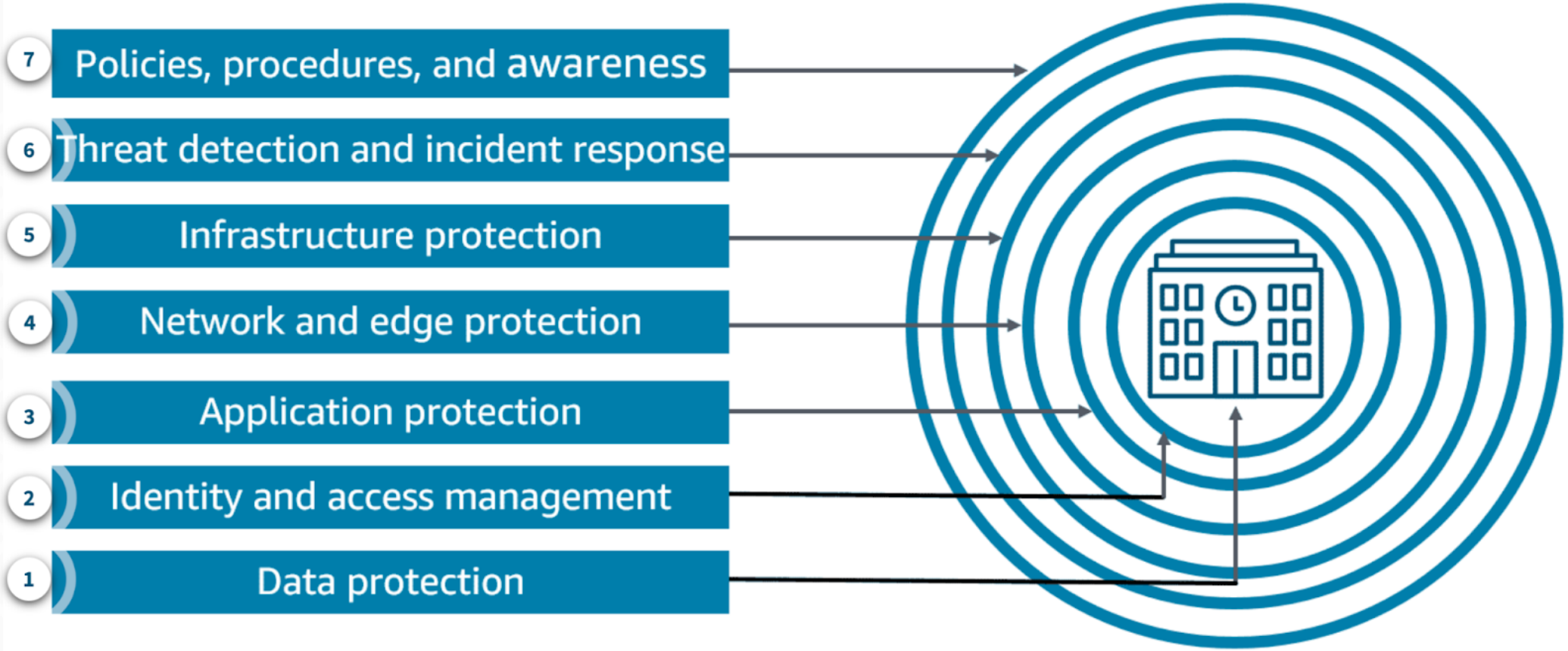
Cumplimiento: Garantizar el cumplimiento normativo de los requisitos en todas las funciones de una organización.

Las organizaciones implementan funciones de seguridad, gobernanza y cumplimiento para garantizar que puedan cumplir con su negocio principal. A veces, los requisitos para estas funciones se denominan los requisitos más importantes, o las cosas que no deben sacrificarse en el desarrollo o la entrega del producto.

CloudOps Guild
Together, towards mastery in Cloud and DevOps

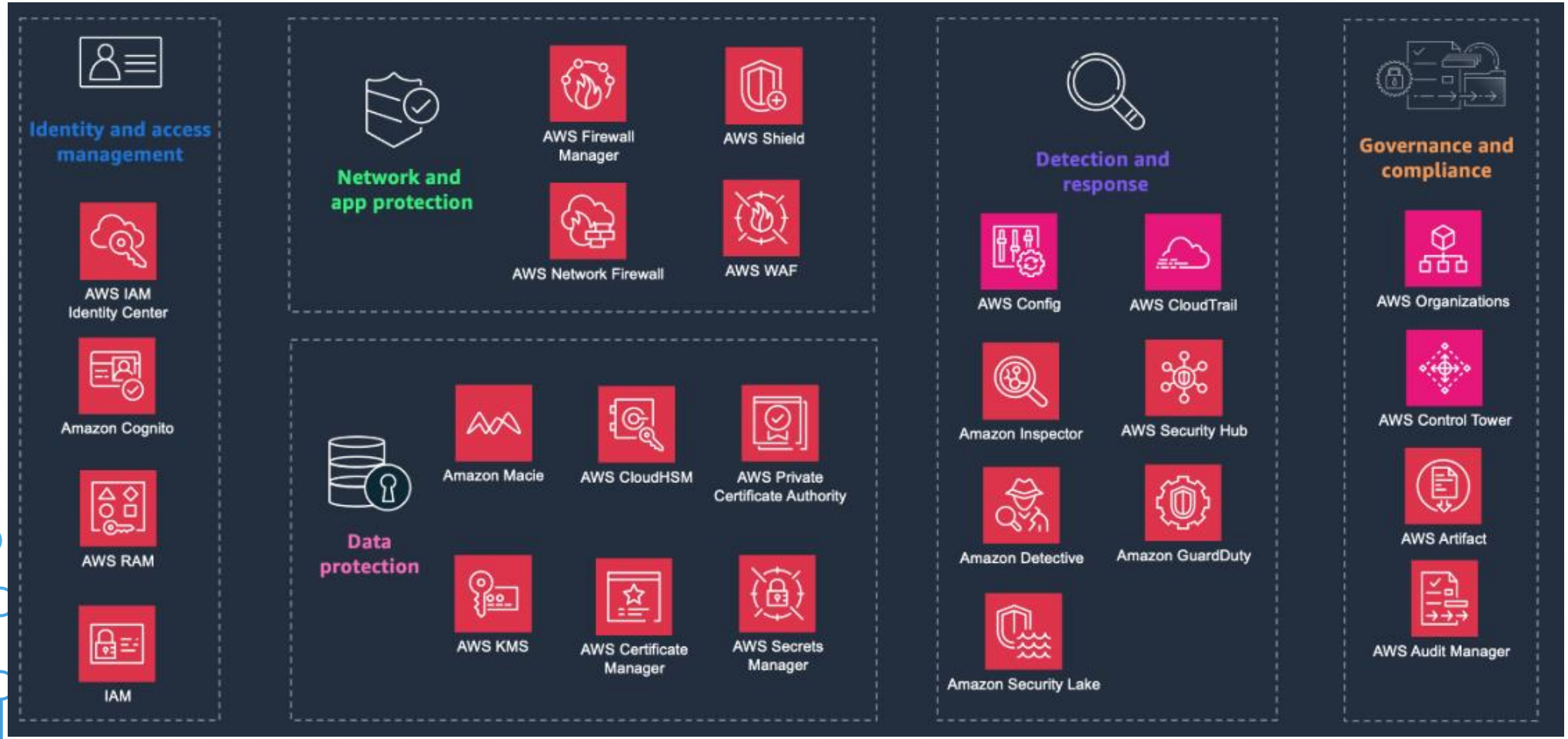
AWS COMMUNITY CARTAGENA

Defensa en profundidad



Servicios de AWS

Seguridad, Identidad y Gobierno



CloudOps Guild
Together, towards mastery in Cloud and DevOps

AWS COMMUNITY CARTAGENA

Recursos adicionales

CloudOps Guild
Together, towards mastery in Cloud and DevOps

https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/

https://aws.amazon.com/es/blogs/machine-learning/architect-defense-in-depth-security-for-generative-ai-applications-using-the-owasp-top-10-for-llms/

https://docs.aws.amazon.com/whitepapers/latest/aws-caf-for-ai/governance-perspective-managing-an-aiml-driven-organization.html

https://aws.amazon.com/es/compliance/

https://aws.amazon.com/es/compliance/resources/

AWS COMMUNITY
CARTAGENA

Agradecimientos

CloudOps Guild
Together, towards mastery in Cloud and DevOps

Próxima sesión:

- Amazon Q
- Amazon Bedrock
- Amazon SageMaker

AWS COMMUNITY
CARTAGENA