UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

"Diseño e implementación de una infraestructura cloud para la gestión documental y búsqueda inteligente de información en instituciones de bomberos"

DIEGO BENJAMÍN ORMEÑO DONOSO

diego.ormeno@usm.cl

Profesor Guía: Gonzalo Mendoza
Profesor Correferente: Rodrigo Pinochet

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Diseño e implementación de una infraestructura cloud para la gestión documental y búsqueda inteligente de información en instituciones de bomberos

Nombre del candidato(a): Diego Benjamín Ormeño Donoso

Carrera / Grado: Ingeniería en Informática

Campus: Viña del Mar **Departamento:** Electrotecnia e informática__

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, ___Gonzalo Mendoza Cárdenas__, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: _____25/05/2026_____ **Firma:** _____

Estudiante o Candidato(a):

Fecha: _____21/05/2026_____ **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Resumen: La Academia Nacional de Bomberos de Chile enfrenta dificultades en la gestión y trazabilidad de la información académica y administrativa asociada a su personal, debido a la dispersión de registros, procesos manuales y falta de centralización. Esta situación dificulta el acceso oportuno a antecedentes relevantes para la gestión formativa y administrativa. Este trabajo propone el diseño e implementación de una plataforma web centralizada para la gestión de perfiles, cursos, certificaciones y documentos asociados, incorporando un asistente virtual que permite realizar consultas en lenguaje natural sobre la información institucional. El objetivo es mejorar la disponibilidad, trazabilidad y accesibilidad de los datos mediante una arquitectura cliente–servidor basada en servicios en la nube. La metodología empleada es de carácter ágil e incremental, con validación funcional a través de un prototipo operativo. Las pruebas realizadas verifican el correcto acceso a la información institucional y la coherencia de las respuestas entregadas por el asistente virtual. Como resultado, se obtiene una solución funcional y escalable, que sienta las bases técnicas para una futura evolución hacia mecanismos avanzados de recuperación semántica de información sobre documentos (enfoque RAG).

Palabras Clave: gestión académica, gestión documental, asistente virtual, computación en la nube, bomberos.

1 Introducción

1.1 Contexto, motivación y problemática:

La Academia Nacional de Bomberos de Chile (ANB) tiene por misión estandarizar la formación y el entrenamiento a nivel nacional, articulando academias locales, instructores certificados y los distintos cuerpos y compañías del país. Para cumplir este objetivo, resulta esencial disponer de información confiable, completa y trazable respecto de las personas (bomberos/as), sus cursos, evaluaciones, certificaciones y vigencias asociadas.

En la práctica, la gestión de dichos antecedentes presenta desafíos relevantes: coexistencia de soportes heterogéneos (documentos físicos, planillas, correos, repositorios aislados), baja estandarización de metadatos y limitada digitalización histórica. Esta situación dificulta el acceso oportuno a la información, incrementa la probabilidad de inconsistencias y retrasa procesos administrativos y académicos (por ejemplo, verificar requisitos de participación, emitir reportes consolidados o comprobar la vigencia de una certificación específica).

El interés de este trabajo surge entonces de la necesidad institucional de modernizar la gestión documental y de información formativa, habilitando trazabilidad y consulta eficiente a escala nacional. Desde la perspectiva de Ingeniería en Informática, el problema permite aplicar conocimientos de arquitecturas en la nube, gestión de datos y técnicas de recuperación de información basadas en inteligencia artificial, con impacto directo en la calidad y oportunidad de los procesos misionales de la ANB.

1.2 Definición del problema:

En el estado actual, los antecedentes administrativos y académicos de bomberos/as se encuentran dispersos y gestionados en gran medida de forma manual, sin un registro maestro unificado por persona ni mecanismos sistemáticos de trazabilidad curricular. Esta fragmentación genera duplicidades, falta de control de versiones y tiempos elevados para realizar búsquedas o validaciones. Adicionalmente, la ausencia de digitalización completa y de metadatos consistentes reduce la capacidad de elaborar reportes y de planificar la formación con base en evidencia.

De persistir este escenario, se proyectan ineficiencias sostenidas en la operación administrativa y académica, pérdida de confiabilidad de la información para la toma de decisiones y mayor riesgo en contextos que demandan verificaciones rápidas (por ejemplo, confirmar competencias habilitantes antes de una actividad operativa). En síntesis, el problema puede enunciarse como: la ANB y los cuerpos de bomberos carecen de un sistema centralizado y estandarizado que registre, organice y trace de manera confiable los antecedentes administrativos y académicos de cada bombero/a, dificultando el acceso oportuno y la gestión a nivel local y nacional.

1.3 Breve descripción sobre la propuesta de solución, los objetivos planteados y el marco teórico adoptado:

El proyecto propone el desarrollo de una plataforma centralizada de gestión administrativa y académica para la Academia Nacional de Bomberos de Chile (ANB), orientada a modernizar la administración de información y fortalecer la trazabilidad institucional. La plataforma busca unificar los registros de formación y certificación del personal bomberil mediante un sistema que integre el registro por bombero, la gestión de cursos y certificaciones, el almacenamiento digital de documentos asociados y un módulo de consulta inteligente en lenguaje natural.
La solución general se estructura como un ecosistema digital donde los datos institucionales y documentos pueden ser cargados, organizados y consultados de forma eficiente, mejorando la disponibilidad de la información para procesos administrativos y académicos (por ejemplo, validación de requisitos, verificación de vigencia de certificaciones y generación de reportes).

- **Objetivo general del proyecto.**

Modernizar la gestión académica y documental de la Academia Nacional de Bomberos de Chile mediante la centralización de los registros institucionales, la digitalización de documentos y la habilitación de mecanismos de consulta eficiente que permitan mejorar la trazabilidad, disponibilidad y confiabilidad de la información formativa y administrativa.

- **Objetivo general de la tesina.**

Diseñar e implementar una infraestructura tecnológica basada en una arquitectura escalable y flujos de procesamiento de información que permitan centralizar la gestión académica y documental institucional, así como habilitar un asistente de consulta en lenguaje natural sobre la información estructurada disponible, validando su factibilidad técnica en un prototipo operativo.

- **Objetivos específicos de la tesina.**

1. Levantar y analizar requerimientos funcionales y no funcionales asociados a la gestión de perfiles, cursos, certificaciones y documentos institucionales.

2. Diseñar una arquitectura cliente–servidor que defina los componentes necesarios para la persistencia de datos estructurados y el almacenamiento de documentos digitales, considerando criterios de seguridad y trazabilidad.

3. Implementar los servicios de aplicación necesarios para la administración de perfiles, cursos, certificaciones y la asociación de documentos institucionales almacenados.

4. Implementar un asistente virtual que permita realizar consultas en lenguaje natural, integrando la obtención de contexto desde la base de datos y la generación de respuestas mediante un modelo de lenguaje.

5. Definir como línea de evolución una integración de mecanismos de recuperación semántica sobre documentos, documentando el flujo técnico y sus consideraciones para una adopción futura.

6. Validar funcionalmente la propuesta mediante un prototipo operativo y casos de prueba representativos, verificando el correcto acceso a la información y la coherencia de las respuestas generadas.

- **En cuanto al marco teórico.**

  Se consideran fundamentos de gestión documental institucional, arquitecturas en la nube y almacenamiento distribuido, diseño de APIs y seguridad en entornos cloud, además de conceptos de búsqueda semántica, modelos generativos y recuperación aumentada de información (RAG) como base conceptual para la evolución del módulo de consulta.

1.4 Breve descripción de la organización del informe en capítulos.

El presente informe se estructura en cinco capítulos, organizados de manera progresiva para reflejar el proceso de diseño, implementación y validación del trabajo de título:

- **Capítulo 1 – Introducción:** presenta el contexto institucional, la definición del problema, la propuesta de solución general, los objetivos del trabajo y la metodología aplicada para su desarrollo y validación.
- **Capítulo 2 – Marco conceptual:** expone los fundamentos teóricos y tecnológicos que sustentan la propuesta, incluyendo conceptos de gestión documental, arquitecturas en la nube, almacenamiento distribuido, búsqueda semántica, modelos generativos y recuperación aumentada de información como línea de evolución del sistema.
- **Capítulo 3 – Diseño de solución:** describe la arquitectura de la plataforma y los componentes desarrollados, diferenciando entre la solución efectivamente implementada —basada en una plataforma web con backend centralizado, base

de datos relacional, almacenamiento cloud y asistente virtual— y una solución avanzada propuesta orientada a la recuperación semántica sobre documentos (enfoque RAG).

- **Capítulo 4 – Validación de la solución:** presenta las pruebas funcionales realizadas sobre el prototipo operativo, los escenarios de uso evaluados y el análisis de resultados en relación con los objetivos definidos, verificando el correcto acceso a la información institucional y la coherencia de las respuestas del asistente virtual.
- **Capítulo 5 – Conclusiones y recomendaciones:** resume los principales aportes del trabajo, identifica sus alcances y limitaciones, y plantea recomendaciones y líneas de trabajo futuro para la evolución de la plataforma.

2 Marco Conceptual / Estado del Arte

2.1 Gestión documental en entornos institucionales

La gestión documental corresponde al conjunto de procesos, técnicas y normas orientadas a controlar, organizar y preservar la información que se genera en una organización durante el desarrollo de sus funciones. Su propósito es asegurar que los documentos —independientemente de su formato o soporte— sean accesibles, auténticos y confiables a lo largo de su ciclo de vida, facilitando la continuidad operativa y la trazabilidad institucional [1].

En el contexto actual de las instituciones públicas y de servicios de emergencia, la gestión documental adquiere un rol estratégico, ya que permite garantizar la disponibilidad oportuna de información crítica, asegurar el cumplimiento normativo y promover la transparencia administrativa.

Cuando estos procesos se ejecutan de forma manual o descentralizada, se incrementa el riesgo de pérdida de información, duplicidad de registros y errores humanos, afectando la eficiencia operativa y la toma de decisiones basada en datos.

Los sistemas modernos de gestión documental se caracterizan por incorporar herramientas de automatización y digitalización, las cuales permiten integrar funciones como la captura de documentos físicos y electrónicos, su almacenamiento seguro en repositorios digitales, la clasificación e indexación mediante metadatos estructurados y la recuperación rápida de la información según distintos criterios, tales como autor, fecha, tipo o contenido.

Además, las tendencias recientes apuntan a la gestión documental inteligente, en la que se aplican técnicas de inteligencia artificial y aprendizaje automático para mejorar la búsqueda de información, el reconocimiento de contenido y la extracción automática de datos relevantes desde documentos estructurados y no estructurados [2][3]. Este enfoque contribuye a reducir la carga administrativa, optimizar los procesos de auditoría y aumentar la trazabilidad dentro de las organizaciones.

En el caso de la Academia Nacional de Bomberos de Chile, la implementación de una gestión documental moderna representa un paso fundamental hacia la consolidación de la información institucional y la optimización de los procesos formativos y administrativos. La digitalización de los registros y la automatización de su gestión

permiten disponer de un historial único por bombero, facilitando la trazabilidad curricular y la toma de decisiones estratégicas a nivel nacional.

2.2 Arquitecturas en la nube y almacenamiento distribuido

Las arquitecturas en la nube representan un modelo tecnológico que permite desplegar, almacenar y operar sistemas informáticos a través de recursos virtualizados, accesibles bajo demanda y con alta escalabilidad. Este paradigma ha sido ampliamente adoptado debido a su capacidad para optimizar el uso de recursos computacionales y reducir la dependencia de infraestructura física local [4].

Este enfoque sustituye la infraestructura tradicional por servicios administrados que ofrecen mayor flexibilidad, disponibilidad y mecanismos de seguridad integrados, factores especialmente relevantes en entornos donde la continuidad operativa y el acceso remoto a la información resultan críticos [5].

En el ámbito de la gestión documental, la nube posibilita centralizar la información institucional, garantizando su accesibilidad, respaldo y control de versiones. Asimismo, los sistemas de almacenamiento distribuido permiten fragmentar y replicar los datos en múltiples nodos, asegurando su integridad y resiliencia ante fallos de hardware o interrupciones del servicio [6].

Para una institución como la Academia Nacional de Bomberos de Chile, el uso de una infraestructura en la nube resulta particularmente adecuado, ya que facilita:

- el acceso simultáneo a la información desde distintas regiones,
- la escalabilidad progresiva frente al crecimiento del volumen documental, y
- la integración con servicios inteligentes, tales como búsquedas avanzadas o análisis automatizado de archivos.

En este sentido, la arquitectura en la nube se consolida como la base tecnológica que sustenta la digitalización, el almacenamiento seguro y la evolución hacia mecanismos avanzados de recuperación de información en sistemas institucionales.

2.3 Seguridad de la información en entornos cloud

La seguridad de la información constituye un aspecto esencial en sistemas institucionales que gestionan datos sensibles. Los principios clásicos de confidencialidad, integridad y disponibilidad (CIA) establecen la base conceptual para garantizar que la información solo sea accesible por entidades autorizadas, no sea alterada de forma indebida y se encuentre disponible cuando sea requerida.

En entornos de computación en la nube, estas prácticas se complementan con estándares y marcos de referencia ampliamente utilizados en la industria, tales como la norma ISO/IEC 27001 para sistemas de gestión de seguridad de la información, las recomendaciones del NIST Cybersecurity Framework y los lineamientos de buenas prácticas en control de accesos e identidad digital.

Estos estándares proporcionan directrices para la definición de políticas de control de acceso, cifrado de datos en tránsito y en reposo, gestión de identidades y auditoría de operaciones. En consecuencia, estos principios y estándares constituyen la base conceptual sobre la cual se diseña el apartado de seguridad de la arquitectura propuesta

en el Capítulo 3, asegurando coherencia entre la fundamentación teórica y la solución técnica implementada.

2.4 Técnicas de búsqueda avanzada y recuperación aumentada (RAG)

La búsqueda avanzada de información constituye un componente fundamental en los sistemas modernos de gestión documental, especialmente en contextos institucionales donde se administran grandes volúmenes de datos heterogéneos. Su objetivo es permitir la recuperación de información relevante considerando no solo coincidencias exactas de palabras clave, sino también el significado y contexto de las consultas realizadas por los usuarios.

Los enfoques tradicionales de búsqueda se basan principalmente en técnicas de coincidencia de palabras clave (*keyword matching*), las cuales presentan limitaciones cuando los términos utilizados en la consulta no coinciden literalmente con el contenido de los documentos. Para superar estas restricciones, los sistemas contemporáneos incorporan técnicas de búsqueda semántica, utilizando representaciones vectoriales del lenguaje (*embeddings*) que permiten medir la similitud semántica entre textos y consultas, mejorando la precisión de los resultados recuperados [7].

En este contexto surge el paradigma de Retrieval-Augmented Generation (RAG), el cual combina motores de recuperación de información con modelos de lenguaje generativos. Este enfoque se estructura, de manera general, en dos etapas principales:

1. **Recuperación de información**, donde un sistema de búsqueda identifica los documentos o fragmentos más relevantes a partir de una consulta en lenguaje natural, utilizando técnicas de similitud semántica.
2. **Generación de respuesta**, donde un modelo de lenguaje procesa la información recuperada y genera una respuesta contextualizada, utilizando los documentos seleccionados como fuente de conocimiento.

La principal ventaja del enfoque RAG es que permite generar respuestas fundamentadas en información específica y controlada, reduciendo la dependencia exclusiva del conocimiento paramétrico del modelo generativo y mejorando la trazabilidad y confiabilidad de las respuestas entregadas [8].

En entornos institucionales, la aplicación de RAG resulta especialmente pertinente para la resolución de consultas complejas sobre información documental, tales como validación de certificaciones, verificación de requisitos o análisis de antecedentes históricos. Sin embargo, su adopción implica desafíos técnicos y operativos asociados a la preparación de los datos, la indexación eficiente de documentos, la gestión de costos y la gobernanza de la información [9].

En el contexto de este trabajo de título, el enfoque RAG se considera como una línea de evolución avanzada para el módulo de consulta inteligente de la plataforma propuesta. La solución implementada valida el uso de consultas en lenguaje natural mediante un asistente virtual apoyado en información estructurada, mientras que la incorporación de recuperación semántica sobre documentos se plantea como una extensión futura, cuyo diseño y factibilidad técnica se analizan dentro del marco conceptual del proyecto.

2.5 Comparativa de soluciones tecnológicas (AWS, Azure, GCP)

En la actualidad, las principales plataformas de servicios en la nube —Amazon Web Services (AWS), Microsoft Azure y Google Cloud Platform (GCP)— ofrecen un conjunto amplio de herramientas orientadas a la gestión de datos, almacenamiento de información y la implementación de soluciones basadas en inteligencia artificial. Estas plataformas comparten principios comunes de escalabilidad, disponibilidad y seguridad, propios de los entornos de computación en la nube. Sin embargo, presentan diferencias relevantes en su grado de integración entre servicios, en la complejidad requerida para configurar flujos completos de procesamiento de información y en su enfoque de interoperabilidad.

Para efectos de este trabajo, la comparación se realiza considerando los siguientes criterios: capacidad de almacenamiento y gestión documental, facilidad de integración entre servicios de almacenamiento, procesamiento de documentos y modelos de lenguaje, soporte para búsqueda semántica, recuperación aumentada (RAG) y modelos generativos, complejidad de orquestación requerida para construir un flujo completo de consulta inteligente, y adecuación al desarrollo de un prototipo funcional con proyección de escalabilidad institucional.

Mientras algunas plataformas priorizan la disponibilidad de un amplio catálogo de servicios desacoplados y altamente configurables, otras enfatizan la integración nativa entre componentes para simplificar el diseño y la operación de soluciones completas. Estas diferencias influyen directamente en el esfuerzo de implementación, mantenimiento y evolución de sistemas institucionales, particularmente en escenarios como el de este proyecto, que requieren combinar almacenamiento documental, procesamiento de información y capacidades de inteligencia artificial para habilitar mecanismos de consultas en lenguaje natural.

Amazon Web Services (AWS)

Amazon Web Services (AWS) se caracteriza por la amplitud y madurez de su catálogo de servicios, incluyendo Amazon S3 para almacenamiento, Amazon Textract para análisis de documentos y Amazon Bedrock para el acceso a modelos generativos, no obstante, la construcción de un flujo completo de gestión documental inteligente en AWS suele requerir la orquestación de múltiples servicios independientes, lo que incrementa la complejidad del diseño y la operación del sistema [15].

Microsoft Azure

Microsoft Azure destaca por su fuerte integración con entornos corporativos y herramientas del ecosistema Microsoft. Sus servicios orientados al análisis documental y búsqueda, como Azure Document Intelligence y Azure AI Search, permiten implementar flujos avanzados de extracción y consulta de información. Sin embargo, la integración con modelos generativos y mecanismos de búsqueda semántica suele implicar la coordinación de múltiples servicios complementarios, aumentando la carga de configuración y operación [16].

Google Cloud Platform (GCP)

Google Cloud Platform (GCP) presenta un enfoque de integración nativa entre servicios de almacenamiento (Cloud Storage), procesamiento de documentos (Document AI) y modelos generativos (Vertex AI). Esta integración facilita la construcción de flujos unificados para la gestión documental y la consulta inteligente de información, reduciendo la necesidad de componentes externos y simplificando la orquestación del sistema [17].

Con el fin de sustentar la selección de la plataforma cloud utilizada en el presente trabajo, se realiza una comparación entre Amazon Web Services (AWS), Microsoft Azure y Google Cloud Platform (GCP).
La evaluación considera tanto aspectos técnicos como económicos, relevantes para el desarrollo de una plataforma institucional de gestión documental con proyección hacia consulta inteligente basada en modelos de lenguaje.

Para ello, se desarrolla una comparación cualitativa basada en criterios de integración tecnológica y una comparación cuantitativa basada en un escenario de referencia de costos.

2.5.1 Criterios de evaluación

La comparación se realiza considerando los siguientes criterios:

1. Capacidad de almacenamiento y gestión documental administrada.
2. Facilidad de integración entre almacenamiento, procesamiento documental y modelos de lenguaje.
3. Soporte para búsqueda semántica y recuperación aumentada (RAG).
4. Complejidad de orquestación arquitectónica para construir el flujo completo.
5. Adecuación al desarrollo de un prototipo institucional escalable.
6. Impacto económico estimado en operación básica y en uso de modelos de lenguaje.

Estos criterios permiten evaluar no solo la disponibilidad de servicios, sino también la factibilidad técnica y económica de implementar la solución propuesta.

2.5.2 Comparativa cualitativa de integración tecnológica

Criterio evaluado	**AWS**	**Azure**	**GCP**
Capacidad de almacenamiento documental administrado	Alta (Amazon S3)	Alta (Azure Blob Storage)	Alta (Cloud Storage)
Integración entre almacenamiento y procesamiento documental	Media (servicios desacoplados)	Media (servicios separados)	Alta (Cloud Storage + Document AI)
Integración con modelos generativos	Media (Amazon Bedrock como servicio separado)	Media (Azure OpenAI Service)	Alta (Vertex AI integrado)
Soporte para búsqueda semántica y RAG	Media–Alta (OpenSearch + Bedrock Knowledge Bases)	Media–Alta (Azure AI Search + Azure OpenAI)	Alta (Vertex AI Search / RAG Engine)

Criterio evaluado	AWS	Azure	GCP
Complejidad de orquestación arquitectónica	Alta	Media	Baja
Adecuación para prototipo institucional escalable	Media	Alta	Alta

Tabla 1. Matriz cualitativa comparativa entre AWS, Azure y GCP.

Fuente: Elaboración propia

La Tabla 1 presenta una matriz cualitativa que relaciona los componentes funcionales requeridos para implementar una solución de búsqueda inteligente sobre documentos institucionales, indicando los servicios equivalentes disponibles en cada proveedor cloud.

A partir de la Tabla 1, se observa que AWS y Azure ofrecen catálogos amplios de servicios, pero requieren una mayor orquestación entre componentes independientes para construir flujos completos de gestión documental inteligente. En contraste, Google Cloud Platform presenta una integración más directa entre servicios de almacenamiento, procesamiento documental y modelos generativos, reduciendo la complejidad de integración para la implementación del prototipo del presente trabajo.

2.5.3 Comparativa cuantitativa de costos de infraestructura base

Con el fin de sustentar objetivamente la selección de la plataforma cloud, se realizó un análisis cuantitativo de costos asociado a la infraestructura base necesaria para la gestión documental del sistema propuesto.

El escenario de evaluación considera una plataforma institucional donde los bomberos pueden cargar, almacenar y descargar documentos digitales tales como certificados de cursos, credenciales y antecedentes administrativos. Para efectos comparativos, se definió un caso de referencia común para los tres proveedores analizados:

- Almacenamiento mensual: 100 GB de documentos en clase estándar/hot.
- Operaciones mensuales: 100.000 operaciones de carga y 100.000 operaciones de lectura.
- Transferencia de entrada: 50 GB/mes (carga de nuevos documentos).
- Transferencia de salida: 300 GB/mes (descarga de documentos por usuarios).
- Región: Sudamérica.

Este escenario representa una carga moderada y realista para una plataforma institucional en etapa inicial, permitiendo comparar los costos base sin incluir aún servicios avanzados de procesamiento o inteligencia artificial.

Las estimaciones se obtuvieron utilizando las calculadoras oficiales de precios de cada proveedor, configuradas bajo parámetros equivalentes. Los resultados se presentan en la Tabla 2.

Los valores corresponden a precios públicos PAYG, sin descuentos por free tier, reservas, ni acuerdos empresariales, moneda USD, región Sudamérica y estimación mensual. No se consideran costos asociados a procesamiento de documentos ni inferencia de modelos de lenguaje, los cuales se analizan posteriormente.

Concepto	AWS (S3 + Data Transfer)	Azure (Blob + Bandwidth)	GCP (Cloud Storage + Data Transfer)
Almacenamiento (100 GB Standard/Hot)	4.81 USD	3.26 USD	3.50 USD
Operaciones (100k cargas + 100k lecturas)	Costo marginal dentro del total	1.47 USD	Costo marginal dentro del total
Transferencia de entrada (50 GB subida)	0.00 USD	0.00 USD	0.00 USD
Transferencia de salida (300 GB descarga)	45.00 USD	36.20 USD	12.00 USD
Costo mensual total estimado	49.81 USD	40.92 USD	15.50 USD

Tabla 2. Comparativa cuantitativa de costos mensuales de infraestructura base.

Fuente: Elaboración propia

A partir de los resultados obtenidos, se observa que, bajo el escenario de referencia propuesto, Google Cloud Platform presenta el menor costo mensual de operación base. Esta diferencia se explica principalmente por el menor costo de transferencia de salida de datos hacia internet en la región sudamericana, así como por la baja tarifa asociada a las operaciones de almacenamiento.

Si bien los tres proveedores ofrecen capacidades equivalentes de almacenamiento documental, el análisis económico evidencia que GCP permite una operación más eficiente en costos para una plataforma institucional en etapa inicial, especialmente en escenarios con alta descarga de documentos por parte de los usuarios finales.

Debe considerarse que estos valores corresponden únicamente a infraestructura base (almacenamiento + operaciones + transferencia). Los costos asociados a procesamiento documental, búsqueda semántica y modelos generativos se analizan en la subsección siguiente.

Este análisis permite establecer una línea base de costos operativos sobre la cual se proyecta la incorporación futura de servicios de procesamiento documental y búsqueda semántica.

2.5.4 Comparativa de servicios para búsqueda inteligente y modelos generativos

Además de la infraestructura base de almacenamiento y transferencia de datos, el sistema propuesto contempla como línea de evolución futura la incorporación de capacidades avanzadas de búsqueda semántica sobre documentos institucionales y generación de respuestas en lenguaje natural, mediante modelos de lenguaje de gran

escala (LLM) y técnicas de recuperación aumentada (Retrieval-Augmented Generation, RAG).

En la implementación validada en este trabajo, el asistente conversacional integrado en la plataforma utiliza Gemini para responder consultas en lenguaje natural a partir de información estructurada recuperada desde la base de datos relacional, sin incorporar aún recuperación semántica directa sobre documentos PDF.

No obstante, dado que uno de los objetivos específicos del proyecto es definir y documentar una línea de evolución hacia recuperación semántica documental, resulta necesario comparar los servicios que cada proveedor cloud ofrece para implementar este tipo de arquitectura en una fase futura del sistema.

En Amazon Web Services, la construcción de un sistema de búsqueda inteligente sobre documentos implica combinar servicios como Amazon Textract para extracción de texto, Amazon OpenSearch para indexación y búsqueda vectorial, y Amazon Bedrock para acceso a modelos generativos. Si bien estos servicios permiten construir una solución equivalente, requieren una orquestación manual significativa entre componentes independientes.

En Microsoft Azure, una arquitectura equivalente considera Azure Document Intelligence para procesamiento documental, Azure AI Search para búsqueda semántica, y Azure OpenAI Service para modelos generativos. Esta integración ofrece capacidades completas, aunque continúa dependiendo de la configuración explícita de múltiples servicios desacoplados.

En Google Cloud Platform, los servicios Cloud Storage, Document AI, Vertex AI Search y Vertex AI Generative Models (Gemini) permiten construir un flujo unificado, donde la ingestión de documentos, indexación semántica y consulta mediante modelos generativos pueden integrarse dentro del ecosistema Vertex AI. Esta integración reduce la necesidad de orquestación externa y simplifica la futura implementación de un sistema RAG institucional.

La Tabla 3 presenta la equivalencia funcional entre proveedores para la implementación de búsqueda semántica y modelos generativos, considerada como extensión futura del sistema.

Componente funcional	**AWS**	**Azure**	**GCP**
Procesamiento de documentos	Amazon Textract	Azure Document Intelligence	Document AI
Almacenamiento de documentos	Amazon S3	Azure Blob Storage	Cloud Storage
Indexación y búsqueda semántica	Amazon OpenSearch (Vector Search)	Azure AI Search	Vertex AI Search
Generación de embeddings	Amazon Bedrock (Titan / Cohere)	Azure OpenAI Embeddings	Vertex AI Embeddings
Modelos generativos (LLM)	Amazon Bedrock	Azure OpenAI Service	Vertex AI Generative Models (Gemini)

Componente funcional	**AWS**	**Azure**	**GCP**
Motor RAG administrado	Requiere orquestación manual	Requiere orquestación manual	Vertex AI RAG Engine (nativo)
Complejidad de integración	Alta	Media	Baja

Tabla 3. Comparativa de servicios para búsqueda inteligente y modelos generativos.

Fuente: Elaboración propia

A partir de la Tabla 3, se observa que AWS y Azure disponen de los componentes necesarios para implementar arquitecturas de búsqueda semántica y modelos generativos, pero requieren la integración manual de múltiples servicios independientes, lo que incrementa la complejidad de diseño y operación del sistema.

En contraste, Google Cloud Platform ofrece una integración más directa entre almacenamiento, procesamiento documental e inteligencia artificial dentro del ecosistema Vertex AI, facilitando una futura evolución hacia recuperación semántica documental sin rediseñar la arquitectura base.

En consecuencia, la selección de Google Cloud Platform no solo responde a criterios de costos de infraestructura base, sino también a la factibilidad técnica de evolución futura, coherente con el alcance real del prototipo validado en este trabajo.

2.6 Conclusiones del marco conceptual y selección del enfoque

A partir del análisis contextual presentado en el Capítulo 1 y de la revisión conceptual desarrollada en este capítulo, se evidencia que los sistemas de gestión documental tradicionales presentan limitaciones significativas en contextos institucionales como el de la Academia Nacional de Bomberos de Chile, principalmente debido a la dispersión de la información, la ausencia de trazabilidad centralizada y la falta de mecanismos de búsqueda eficientes sobre grandes volúmenes de datos.

La revisión del estado del arte permitió identificar que las soluciones basadas en arquitecturas en la nube ofrecen una respuesta efectiva frente a estos desafíos, al proporcionar escalabilidad, alta disponibilidad y facilidad de integración entre servicios. Asimismo, los enfoques de búsqueda semántica y recuperación aumentada de información (RAG) representan una línea tecnológica relevante para la evolución futura del sistema, especialmente para mejorar el acceso y comprensión de información documental en escenarios institucionales complejos.

Este marco conceptual respalda directamente los objetivos específicos del presente trabajo de título, en particular aquellos orientados al diseño de una arquitectura cloud, a la centralización de la información institucional, y a la implementación de un módulo de consulta en lenguaje natural que permita acceder de forma eficiente a los datos académicos y administrativos. En este sentido, el énfasis del proyecto se centra en la validación técnica de una solución funcional, dejando la incorporación de recuperación semántica avanzada sobre documentos como una extensión futura del sistema.

La comparativa entre los principales proveedores cloud (GCP, AWS y Azure) permitió determinar que Google Cloud Platform ofrece un conjunto de herramientas coherente con los requerimientos del proyecto, destacando por la integración nativa entre servicios de almacenamiento, procesamiento de documentos y modelos de inteligencia artificial. Esta integración resulta especialmente relevante para cumplir los objetivos de simplicidad arquitectónica, escalabilidad y mantenibilidad definidos en la tesina.

En consecuencia, se selecciona Google Cloud Platform como entorno principal de desarrollo e implementación de la solución propuesta, por su equilibrio entre facilidad de integración, soporte nativo para servicios de inteligencia artificial y adecuación al alcance real del proyecto. En particular, se reconoce el potencial de herramientas como Vertex AI para una futura evolución hacia mecanismos avanzados de recuperación aumentada, sin que ello constituya el núcleo de la implementación validada en este trabajo de título.

3 Propuesta de Solución

3.1 Metodología de desarrollo de la solución

El desarrollo del proyecto se estructura bajo un enfoque ágil e incremental, inspirado en la metodología *Scrum*, con el propósito de garantizar un avance continuo, flexible y orientado a resultados verificables. Este enfoque permite dividir el proceso en iteraciones o sprints, en los que se planifican, diseñan, implementan y prueban funcionalidades específicas del sistema, priorizando aquellas que aportan mayor valor institucional.

Cada iteración contempla tres fases principales:

1. **Análisis y diseño**, donde se identifican requerimientos técnicos, se definen componentes arquitectónicos y se seleccionan las tecnologías necesarias.
2. **Implementación**, orientada al desarrollo modular de los flujos de carga, almacenamiento y consulta sobre información estructurada, incluyendo la integración del asistente conversacional.
3. **Pruebas funcionales iniciales**, mediante la ejecución de pruebas sobre el prototipo operativo para comprobar el cumplimiento de los objetivos definidos y el correcto funcionamiento del módulo de consulta.

Adicionalmente, la metodología contempla una fase de definición y exploración técnica para la integración futura de mecanismos de recuperación semántica sobre documentos, en la cual se analiza el flujo de indexación, recuperación y generación de respuestas, sin comprometer su despliegue productivo dentro del alcance principal del prototipo validado.

El carácter incremental de la metodología permite ajustar la solución de acuerdo con la retroalimentación obtenida durante las revisiones parciales, asegurando que el resultado final cumpla tanto con las necesidades institucionales como con los estándares técnicos esperados en un entorno de producción.

3.2 Descripción general de la solución

La solución propuesta tiene como objetivo el desarrollo de un sistema de gestión documental y administrativa orientado a la centralización, trazabilidad y consulta eficiente de información institucional, asociado a los integrantes de la Academia Nacional de Bomberos de Chile.

El sistema busca reemplazar los procesos manuales actuales, caracterizados por la dispersión de la información y la falta de control sobre los antecedentes académicos y administrativos, mediante una infraestructura cloud centralizada y automatizada que soporta el registro, validación y consulta estructurada de datos y documentos institucionales a través de una plataforma web.

Delimitación del alcance de la solución.

Con el objetivo de mantener consistencia entre el diseño propuesto y la implementación efectivamente validada, se distinguen dos alcances. En primer lugar, la implementación actual corresponde a una plataforma centralizada para la gestión de información institucional y documentos asociados, e integra un asistente conversacional que utiliza un modelo generativo para responder consultas en lenguaje natural a partir de contexto construido desde información estructurada almacenada en la base de datos relacional. En segundo lugar, se define como línea de evolución la incorporación de mecanismos de búsqueda semántica sobre documentos PDF mediante un enfoque RAG, lo cual permitiría recuperar fragmentos relevantes desde contenido documental y utilizarlos como sustento de respuestas.

Por tanto, el flujo validado en este trabajo se limita a:

Carga → Almacenamiento → Consulta inteligente (datos estructurados) → Visualización

Dejando la indexación y recuperación semántica de documentos como extensión futura.

El diseño de la solución contempla una arquitectura modular basada en servicios en la nube, la cual permite gestionar el ciclo de información institucional de forma escalable y segura, considerando tanto la implementación actual como la evolución futura propuesta.

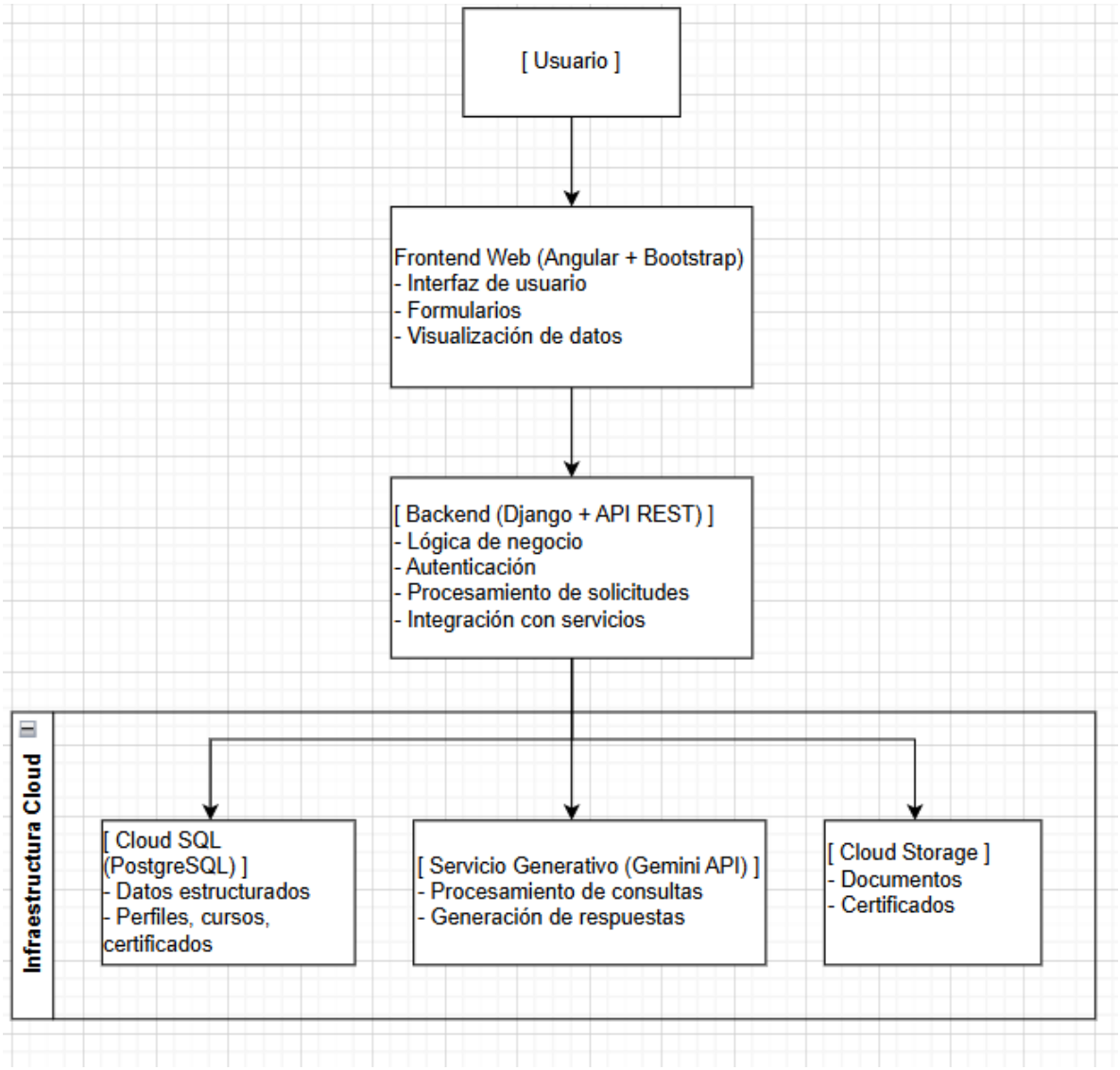


Figura 1. Arquitectura de la solución basada en servicios cloud.

Fuente: Elaboración propia

Desde el punto de vista funcional, la plataforma integra un asistente virtual de consulta, que permite realizar preguntas en lenguaje natural sobre la información estructurada almacenada en la base de datos. El backend del sistema interpreta las consultas, obtiene el contexto relevante desde la base de datos y lo utiliza como entrada para la generación de respuestas mediante un modelo de lenguaje, validando así la factibilidad técnica de incorporar capacidades de consulta inteligente en el sistema.

A continuación, se presenta un diagrama de flujo que describe el proceso de interacción del usuario con la plataforma y el procesamiento de la información dentro del sistema.

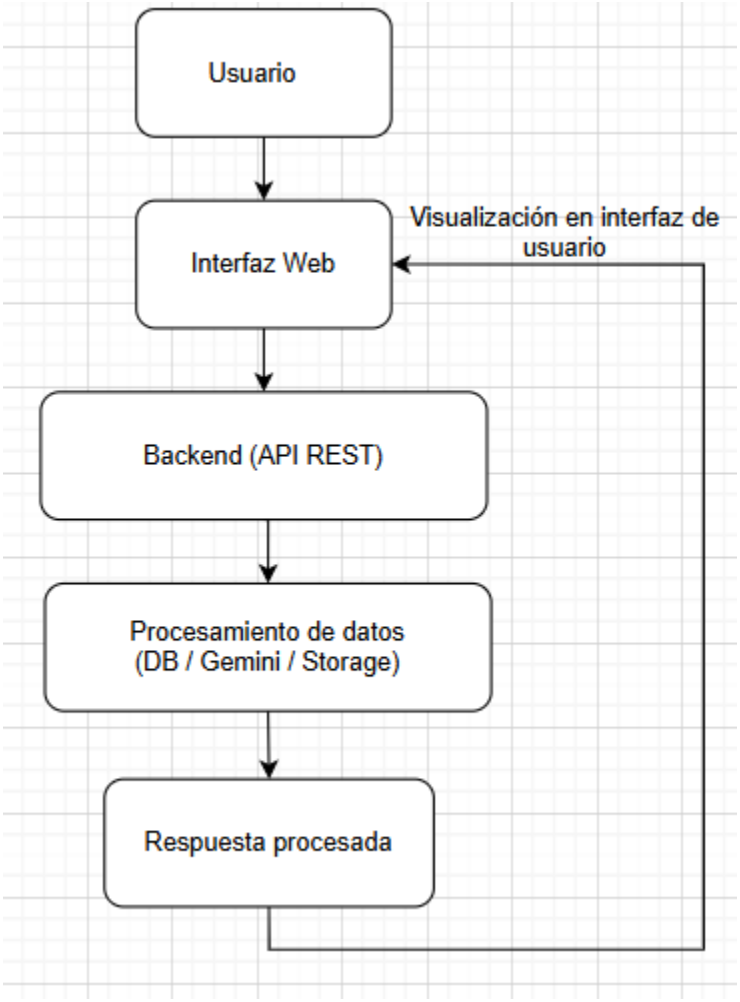


Figura 2. Flujo de procesamiento de la información en la plataforma.

Fuente: Elaboración propia

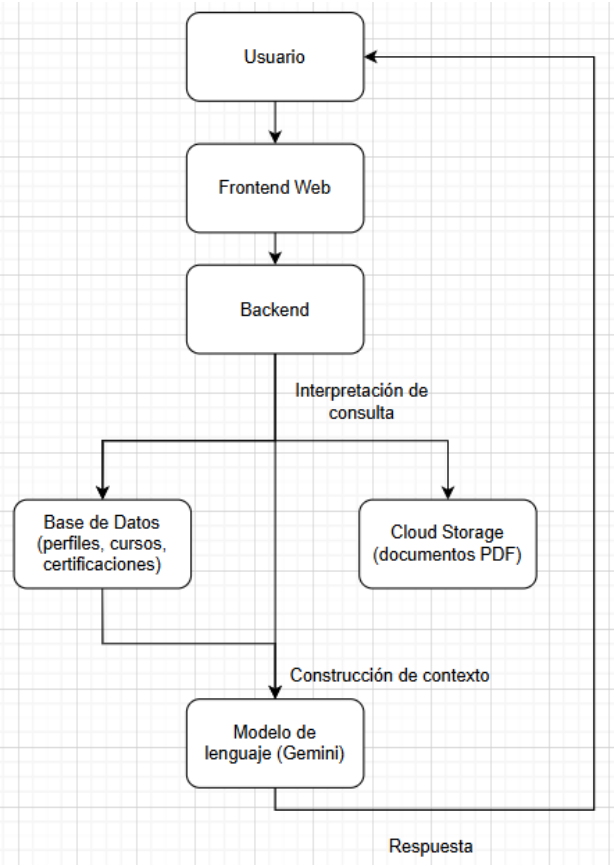


Figura 3**.** Flujo general de procesamiento y consulta de información en la plataforma (implementación actual).

Fuente: Elaboración propia

Cabe destacar que, en la implementación validada, el modelo generativo no accede directamente a los documentos PDF, sino que opera únicamente con contexto textual controlado construido por el backend a partir de información estructurada.

Finalmente, el sistema se concibe como una base tecnológica escalable que permita futuras ampliaciones, tales como módulos de gestión académica, control de certificaciones o análisis histórico de información, acompañando el proceso de transformación digital de la Academia Nacional de Bomberos de Chile.

3.3 Arquitectura propuesta

Esta sección presenta la arquitectura desde una perspectiva lógica y conceptual, mientras que los aspectos relacionados con infraestructura y despliegue se detallan en la sección 3.4.

La arquitectura de la solución se diseñó bajo un enfoque cloud-native, priorizando modularidad, escalabilidad y automatización de los procesos asociados a la gestión documental institucional. La solución se implementa sobre Google Cloud Platform (GCP), utilizando servicios administrados que permiten desacoplar responsabilidades y facilitar la integración entre los distintos componentes del sistema.

Con el fin de mantener coherencia entre el diseño propuesto y la implementación efectivamente validada, se distinguen dos alcances arquitectónicos:

- **Implementación actual validada**, correspondiente al prototipo funcional desarrollado y probado en este trabajo.
- **Línea de evolución futura**, correspondiente a la incorporación de mecanismos de recuperación semántica documental mediante un enfoque RAG.

Las siguientes subsecciones presentan ambas arquitecturas de manera diferenciada.

3.3.1 Arquitectura de la implementación actual validada

En la implementación actualmente validada, el sistema opera bajo un esquema cliente–servidor, donde el objetivo principal es centralizar la gestión de información institucional estructurada y habilitar un asistente virtual de consulta en lenguaje natural.

La plataforma se organiza siguiendo una arquitectura por capas, donde cada nivel cumple una responsabilidad específica dentro del funcionamiento global del sistema. Esta separación facilita la mantenibilidad, escalabilidad y evolución futura de la plataforma.

La Figura 4 presenta la arquitectura general organizada por capas y los principales componentes tecnológicos utilizados.

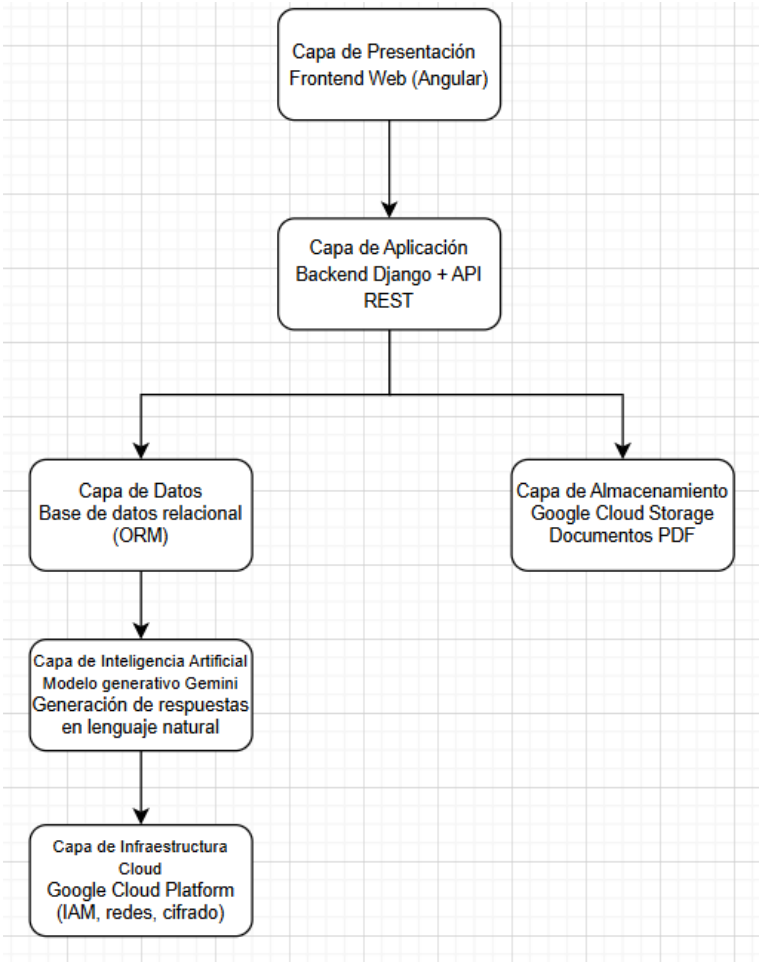


Figura 4. Arquitectura general de la solución organizada por capas.

Fuente: Elaboración propia

- **Capa de presentación:** Frontend web desarrollado en Angular, encargado de la interacción con usuarios institucionales.
- **Capa de aplicación:** Backend desarrollado en Django + API REST, que concentra la lógica de negocio, validación de operaciones y orquestación entre servicios cloud.
- **Capa de datos:** Base de datos relacional PostgreSQL (Cloud SQL en GCP), accedida desde el backend mediante Django ORM, donde se almacena la información estructurada institucional (perfiles, cursos, certificaciones).
- **Capa de almacenamiento:** Google Cloud Storage, donde se almacenan los documentos institucionales en formato PDF.
- **Capa de inteligencia artificial:** Modelo generativo Gemini, que recibe un contexto textual construido por el backend y genera respuestas en lenguaje natural.
- **Capa de infraestructura:** Google Cloud Platform, que provee servicios de seguridad, cifrado, redes y control de accesos mediante IAM.

En esta arquitectura, el modelo generativo no accede directamente a los documentos PDF, sino que recibe únicamente un contexto textual controlado construido por el backend a partir de consultas estructuradas a la base de datos relacional. Esto permite validar la factibilidad técnica de incorporar un asistente inteligente sin exponer directamente los documentos institucionales al modelo.
El flujo funcional de alto nivel de la implementación actual es:

Carga → Almacenamiento → Consulta inteligente (datos estructurados) → Presentación de resultados

Este flujo es el que fue implementado y validado funcionalmente en el prototipo desarrollado.

3.3.2 Arquitectura de evolución futura hacia recuperación semántica documental (RAG)

Si bien la implementación actual valida la consulta inteligente sobre información estructurada, uno de los objetivos específicos del proyecto es definir una línea de evolución futura hacia mecanismos de recuperación semántica sobre documentos PDF institucionales.

En esta arquitectura futura, los documentos almacenados en Google Cloud Storage serían procesados mediante extracción de texto (OCR/parsing), generación de representaciones vectoriales (embeddings) y almacenamiento en un corpus semántico. Ante una consulta en lenguaje natural, el sistema recuperaría fragmentos relevantes desde dicho corpus y los utilizaría como contexto de entrada para el modelo generativo, siguiendo un enfoque Retrieval-Augmented Generation (RAG).

La Figura 5 presenta el flujo propuesto de evolución hacia recuperación semántica documental.

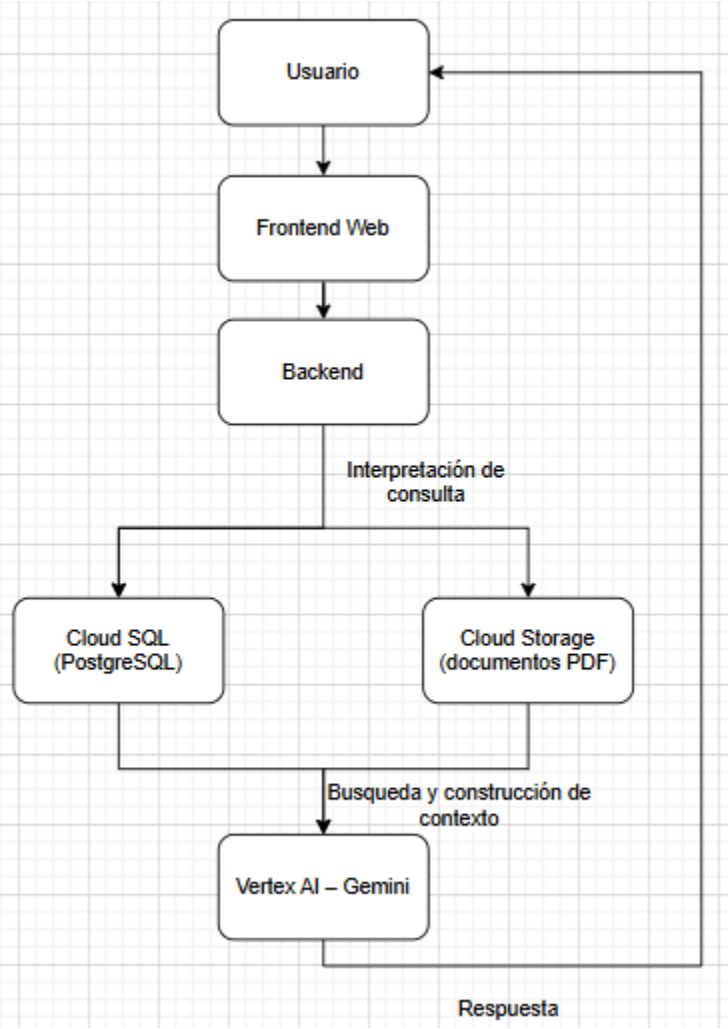


Figura 5. Flujo propuesto de evolución hacia recuperación semántica documental (RAG).

Fuente: Elaboración propia

En esta línea de evolución:

1. Los documentos PDF almacenados en Cloud Storage son procesados para extraer texto.
2. El texto se fragmenta y transforma en embeddings.
3. Los embeddings se almacenan en un motor de búsqueda semántica.
4. Ante una consulta, se recuperan fragmentos relevantes.
5. El modelo generativo Gemini utiliza estos fragmentos como contexto para generar respuestas fundamentadas en contenido documental.

Es importante destacar que esta arquitectura RAG no forma parte de la implementación actualmente validada, sino que se documenta como una extensión futura coherente con los objetivos de evolución tecnológica del sistema.

3.4 Componentes principales de la solución (implementación actual validada)

La solución implementada y validada en este trabajo corresponde a una arquitectura orientada a la gestión documental institucional y a la consulta inteligente de información estructurada, construida bajo un esquema cliente-servidor e integrada con servicios cloud de Google Cloud Platform. Desde el punto de vista de infraestructura, el sistema combina una capa de presentación web, una capa de aplicación, un componente de persistencia de datos estructurados, un componente de almacenamiento documental y un componente de generación de respuestas mediante inteligencia artificial.

A diferencia de una descripción puramente funcional, en esta sección se explicitan los principales componentes de la solución implementada, su ubicación lógica dentro de la arquitectura y la relación que mantienen con los servicios cloud utilizados. Esto permite evidenciar cómo la infraestructura propuesta soporta el flujo de carga, almacenamiento, consulta y recuperación de información institucional.

La figura 6 presenta la arquitectura lógica general de la solución implementada, identificando la interacción entre el frontend, el backend y los servicios cloud utilizados para persistencia, almacenamiento documental y consulta inteligente.

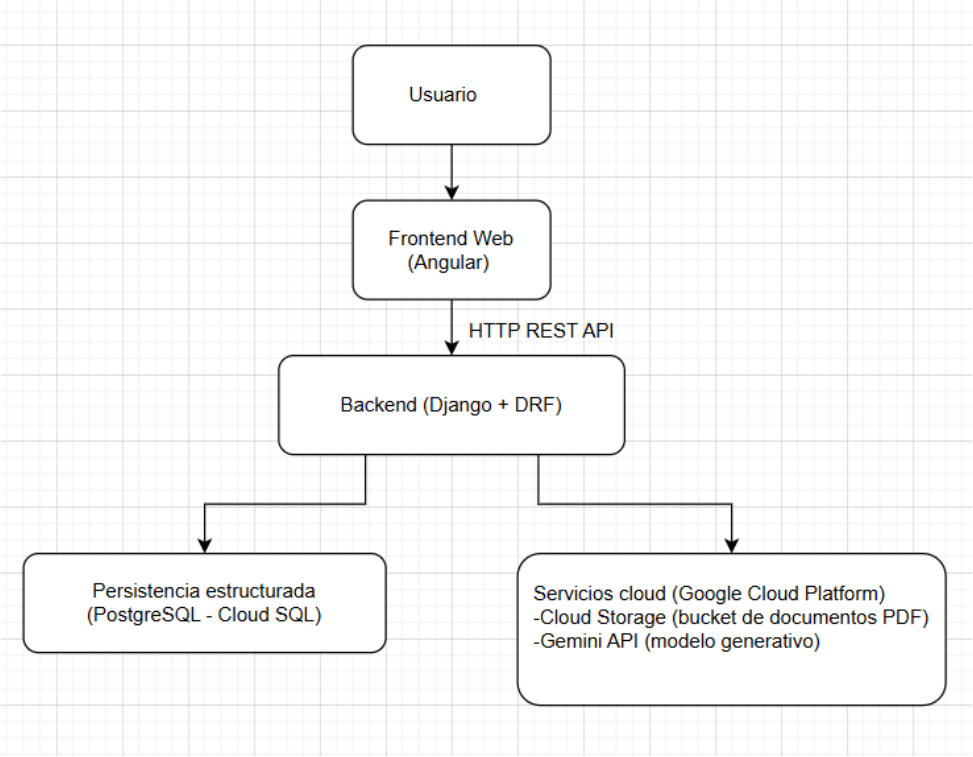


Figura 6. Arquitectura lógica de la solución implementada.

Fuente: Elaboración propia

3.4.1 Diseño de infraestructura y despliegue de la solución

A diferencia de la arquitectura lógica presentada en la sección 3.3, esta subsección se enfoca específicamente en el diseño de infraestructura cloud y en la forma en que los componentes del sistema se organizan desde el punto de vista de despliegue, almacenamiento, segmentación y disponibilidad.

Desde el punto de vista de infraestructura, la arquitectura se concibe bajo un enfoque distribuido sobre servicios cloud, donde los componentes del sistema se organizan en distintos segmentos lógicos según su responsabilidad dentro del flujo de procesamiento de información.

3.4.1.1 Segmentación lógica de la arquitectura

A nivel conceptual, la solución distingue tres segmentos principales:

1. Capa de acceso (cliente)
Corresponde a la aplicación frontend desarrollada en Angular, la cual es consumida por los usuarios mediante un navegador web. Esta capa actúa como punto de entrada al sistema y se comunica exclusivamente con el backend a través de API REST.

2. Capa de aplicación (backend)
Corresponde al servicio backend implementado en Django + Django REST Framework, el cual centraliza la lógica de negocio, validación de datos, autenticación y orquestación de servicios.
Este componente actúa como intermediario entre la capa de acceso y los servicios cloud, evitando accesos directos desde el cliente hacia la base de datos o el almacenamiento documental.

3. Capa de servicios cloud
Incluye los servicios administrados utilizados por la solución:

- Base de datos relacional (PostgreSQL / Cloud SQL): utilizada para almacenar información estructurada institucional (perfiles, cursos, certificaciones y metadatos).
- Google Cloud Storage (buckets): utilizado como repositorio de almacenamiento de documentos PDF institucionales.
- Servicios de inteligencia artificial (Gemini): utilizados para la generación de respuestas en lenguaje natural a partir de contexto construido por el backend.

3.4.1.2 Diseño de red y segmentación

Desde el punto de vista de infraestructura cloud, el diseño de la solución es compatible con un esquema basado en redes virtuales (VPC), donde los componentes pueden organizarse en segmentos con distintos niveles de exposición.

En un escenario de despliegue productivo, la arquitectura considera:

- Una zona de acceso público, donde se expone la aplicación web para los usuarios.
- Una zona de aplicación, donde reside el backend, encargado de procesar las solicitudes.

- Una zona de servicios internos, donde se ubican la base de datos y el almacenamiento, accesibles únicamente desde el backend.

Este enfoque permite aislar los componentes críticos del sistema, mejorar la seguridad y controlar el acceso a los recursos.

Es importante destacar que esta segmentación corresponde a un diseño teórico de infraestructura, ya que la implementación validada en el prototipo se realizó utilizando servicios administrados sin configuración explícita de redes virtuales.

Cabe señalar que, en un escenario productivo, este diseño podría materializarse mediante la configuración de una Virtual Private Cloud (VPC), permitiendo definir subredes, políticas de acceso y segmentación de tráfico entre los distintos componentes del sistema. Sin embargo, dado el uso de servicios administrados en el prototipo, estos aspectos son abstraídos por la plataforma, manteniendo estándares de seguridad y aislamiento sin requerir configuración explícita.

3.4.1.3 Gestión del almacenamiento

El almacenamiento de la solución se encuentra desacoplado en dos niveles claramente diferenciados:

- Los datos estructurados se almacenan en la base de datos relacional (PostgreSQL / Cloud SQL).
- Los archivos binarios (documentos PDF) se almacenan en buckets de Google Cloud Storage.

El acceso a ambos sistemas es gestionado exclusivamente por el backend, el cual coordina:

- la carga de archivos hacia el bucket,
- la recuperación de documentos,
- y la asociación entre archivos y registros estructurados.

Este enfoque permite mejorar la escalabilidad, la seguridad y la mantenibilidad del sistema, evitando dependencias directas entre la base de datos y los archivos almacenados.

En particular, el uso de Google Cloud Storage como almacenamiento de objetos permite manejar grandes volúmenes de documentos de manera eficiente, aprovechando mecanismos de redundancia, durabilidad y disponibilidad propios de la infraestructura distribuida del proveedor.

3.4.1.4 Disponibilidad y escalabilidad

En la implementación validada, la solución corresponde a un prototipo funcional, por lo que no se incorporan mecanismos avanzados de alta disponibilidad a nivel de aplicación, tales como balanceadores de carga o despliegues multi-zona del backend.

Sin embargo, la arquitectura se apoya en servicios administrados de Google Cloud Platform, los cuales proporcionan:

- alta disponibilidad a nivel de infraestructura,
- redundancia en almacenamiento (Cloud Storage),
- y soporte para despliegues escalables en servicios como Cloud SQL.

Adicionalmente, la arquitectura propuesta es compatible con la incorporación de mecanismos avanzados de alta disponibilidad, tales como balanceadores de carga (Cloud Load Balancing), despliegues multi-zona del backend y replicación de la base de datos, sin requerir modificaciones estructurales en el diseño actual.

3.4.2 Esquema general de la arquitectura

La arquitectura implementada se organiza en los siguientes componentes principales:

- Capa de acceso y presentación, correspondiente a la aplicación web desarrollada en Angular.
- Capa de aplicación, correspondiente al backend desarrollado en Django y Django REST Framework.
- Componente de persistencia estructurada, correspondiente a una base de datos relacional compatible con PostgreSQL, considerando Cloud SQL como servicio de despliegue cloud.
- Componente de almacenamiento documental, correspondiente a Google Cloud Storage para el resguardo de documentos institucionales.
- Componente de consulta inteligente, correspondiente a la integración con Gemini para la generación de respuestas en lenguaje natural.

Estos componentes se comunican a través de interfaces definidas, principalmente mediante API REST, manteniendo separación de responsabilidades y desacoplamiento entre presentación, lógica de aplicación, almacenamiento y servicios de inteligencia artificial.

3.4.3 Componente de interfaz web (Frontend)

La capa de presentación está implementada mediante una aplicación web desarrollada en Angular, a través de la cual los usuarios institucionales interactúan con el sistema. Desde esta interfaz se ejecutan las principales operaciones funcionales del prototipo, incluyendo carga de documentos, gestión de certificaciones, navegación por perfiles y uso del asistente conversacional.

Este componente no accede directamente a la base de datos ni a los servicios cloud, sino que interactúa exclusivamente con el backend mediante solicitudes HTTP a la API REST. Esta separación permite mantener un modelo de arquitectura desacoplada, donde la lógica de validación, persistencia y orquestación se concentra en la capa de aplicación.

3.4.4 Componente backend y lógica de aplicación

El backend de la solución está desarrollado en Django, utilizando Django REST Framework para la exposición de endpoints y la gestión de operaciones sobre los recursos del sistema. Este componente actúa como núcleo de la arquitectura, centralizando la lógica de negocio y coordinando la interacción entre la interfaz web, la persistencia estructurada, el almacenamiento documental y el asistente conversacional.

Entre sus responsabilidades principales se encuentran:

- Recepción de solicitudes provenientes del frontend.
- Validación de archivos y metadatos asociados.
- Persistencia de información estructurada en la base de datos.
- Coordinación del almacenamiento de archivos en Google Cloud Storage.
- Recuperación de datos para consultas institucionales.
- Construcción de contexto textual controlado para el módulo de consulta inteligente.
- Integración con Gemini para la generación de respuestas.

En términos de diseño, el backend constituye el punto de control principal del sistema, garantizando consistencia operativa, separación de responsabilidades y trazabilidad entre los distintos componentes de la solución.

3.4.5 Componente de persistencia estructurada

La solución contempla una capa de persistencia estructurada destinada al almacenamiento de perfiles de bomberos, cursos, certificaciones, relaciones académicas y metadatos documentales. Desde el punto de vista arquitectónico, la propuesta considera el uso de una base de datos relacional PostgreSQL desplegada mediante el servicio administrado Cloud SQL de Google Cloud Platform, lo que permite proyectar el sistema hacia un entorno cloud escalable y mantenible.

En la validación funcional del prototipo, esta persistencia se ejecutó en un entorno de desarrollo controlado, manteniendo compatibilidad con una migración posterior hacia PostgreSQL/Cloud SQL sin modificar la estructura lógica del sistema. De esta forma, la arquitectura implementada y el modelo de datos permanecen alineados con el despliegue cloud objetivo.

La persistencia estructurada se utiliza exclusivamente para información tabular y relacional, diferenciándose del almacenamiento de archivos binarios, lo que facilita la organización de la información, la trazabilidad institucional y la evolución futura del sistema.

3.4.6 Componente de almacenamiento documental en Google Cloud Storage

Los documentos institucionales, tales como certificados y archivos PDF asociados a cursos o perfiles, se almacenan en un bucket de Google Cloud Storage, el cual actúa como repositorio de objetos binarios dentro de la arquitectura cloud.

El proceso de carga documental es gestionado por el backend, que recibe el archivo desde la interfaz web, valida su formato y tamaño, y posteriormente coordina su almacenamiento en el bucket correspondiente. En paralelo, los metadatos y relaciones del documento son registrados en la capa de persistencia estructurada, permitiendo asociar cada archivo con su contexto institucional.

Este enfoque desacopla el almacenamiento de documentos del almacenamiento de metadatos, lo que mejora la escalabilidad, facilita la administración de archivos y reduce la dependencia entre la base de datos y los objetos binarios.

La Figura 7 presenta el flujo del módulo de carga documental, mostrando la interacción entre el frontend, el backend, la validación del archivo y su posterior registro en la persistencia estructurada y en el almacenamiento cloud.

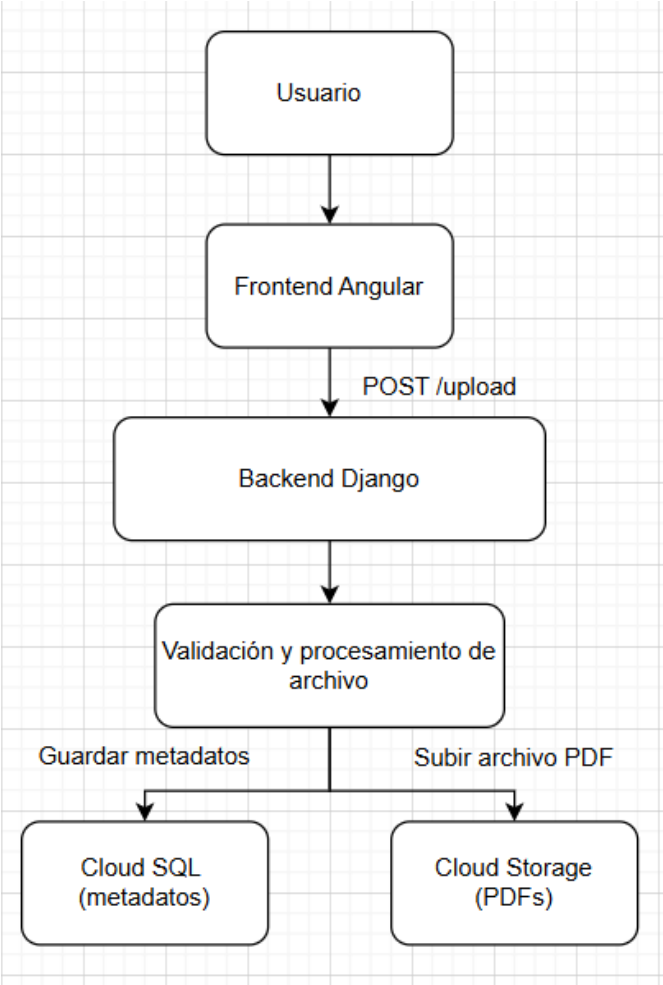


Figura 7. Flujo de carga documental.

Fuente: Elaboración propia

3.4.7 Componente de asistente conversacional con Gemini

La solución incorpora un componente de consulta inteligente basado en Gemini, el cual permite responder preguntas en lenguaje natural sobre la información estructurada del sistema. En la implementación actual validada, el asistente no realiza recuperación

semántica directa sobre documentos PDF, sino que opera a partir de datos estructurados obtenidos por el backend.

El flujo general de operación es el siguiente:

1. El usuario realiza una consulta desde la interfaz web.
2. El backend recibe la solicitud y consulta los datos estructurados pertinentes.
3. A partir de esos datos, el sistema construye un contexto controlado.
4. El contexto es enviado al modelo Gemini.
5. El modelo genera una respuesta en lenguaje natural.
6. La respuesta es retornada al frontend para su visualización.

Este diseño permite incorporar capacidades de consulta inteligente sin exponer directamente los documentos institucionales al modelo generativo y sin requerir aún una arquitectura RAG completa.

La Figura 8 presenta el flujo del asistente conversacional implementado, destacando el rol del backend en la consulta de datos estructurados, la construcción de contexto controlado y la interacción con Gemini.

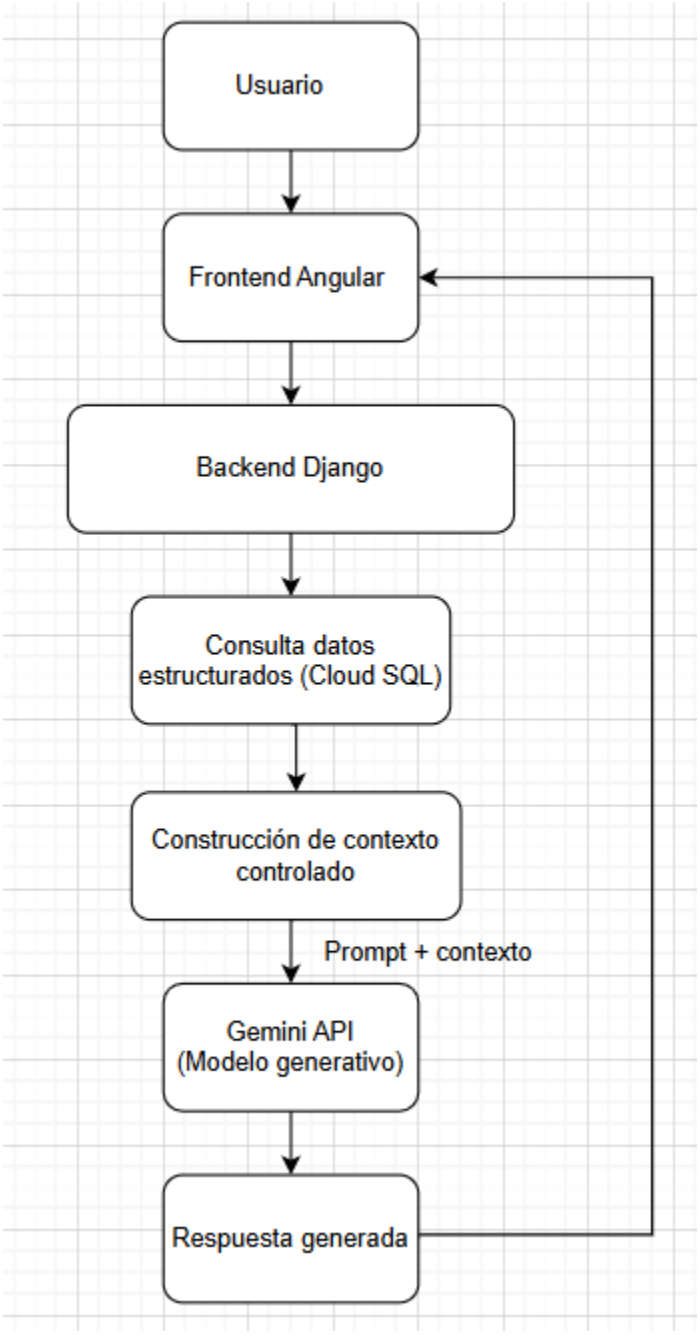


Figura 8. Flujo de consulta del asistente conversacional.

Fuente: Elaboración propia

3.4.8 Seguridad, aislamiento y disponibilidad

La seguridad del sistema se aborda tanto a nivel de aplicación como de infraestructura. En la capa de aplicación, el backend implementa autenticación, control de acceso basado

en roles y validación de operaciones, asegurando que cada usuario acceda únicamente a la información habilitada para su perfil.

En la capa de infraestructura, los servicios cloud utilizados incorporan mecanismos de seguridad propios de Google Cloud Platform, incluyendo cifrado en tránsito y en reposo, además de control de acceso mediante políticas IAM en los servicios correspondientes.

Desde el punto de vista de disponibilidad, la solución se apoya en servicios administrados de la nube, particularmente para almacenamiento documental y servicios de plataforma. No obstante, dado que la implementación validada corresponde a un prototipo funcional, no se incorporan todavía mecanismos avanzados de alta disponibilidad extremo a extremo, tales como balanceo de carga dedicado, despliegue multi-zona de la capa de aplicación o replicación multi-región.

3.4.9 Línea de evolución futura: búsqueda semántica sobre documentos (RAG)

Como línea de evolución tecnológica del sistema, se propone incorporar en el futuro un mecanismo de recuperación semántica sobre documentos institucionales PDF mediante un enfoque Retrieval-Augmented Generation (RAG).

En este escenario, los documentos almacenados en Google Cloud Storage serían procesados para extraer su contenido textual, segmentarlo en fragmentos, generar embeddings y almacenar dichos vectores en un repositorio semántico. Frente a una consulta en lenguaje natural, el sistema recuperaría fragmentos relevantes y los utilizaría como contexto para el modelo generativo.

Esta arquitectura permitiría que Gemini respondiera no solo a partir de información estructurada, sino también utilizando contenido documental institucional como sustento de la respuesta. Sin embargo, esta capacidad no forma parte de la implementación actualmente validada, sino que se documenta como una evolución futura coherente con los objetivos de la solución.

3.5 Rol personal en la implementación

La participación del autor en el proyecto se enfocó en el diseño, configuración e implementación de la infraestructura cloud y del flujo de procesamiento de información que fue efectivamente desarrollado y validado en esta tesina.

El trabajo se enmarca en el ámbito de la gestión de datos y servicios en la nube, integrando componentes de infraestructura cloud con herramientas de inteligencia artificial para habilitar la gestión documental y la consulta inteligente sobre información institucional estructurada.

Desde el alcance de este trabajo de título, el énfasis estuvo puesto en:

- El diseño de la arquitectura cloud que soporta la plataforma.
- La orquestación entre frontend, backend y servicios cloud.
- La implementación del flujo de carga, almacenamiento y consulta sobre información estructurada.
- La integración del asistente conversacional basado en Gemini.

- La definición de la línea de evolución futura hacia búsqueda semántica documental (RAG), sin que esta haya sido implementada en el prototipo validado.

A continuación, se describen las principales etapas del trabajo realizado, indicando explícitamente los componentes cloud involucrados.

a) Diseño del flujo cloud

En esta etapa se realizó el diseño lógico del flujo de procesamiento de información institucional, definiendo los componentes necesarios para cubrir las fases implementadas:

Carga → Almacenamiento → Consulta inteligente → Visualización.

El diseño estableció:

- El rol del backend Django como capa de orquestación.
- La base de datos relacional PostgreSQL (Cloud SQL) como repositorio de información estructurada.
- El almacenamiento en Google Cloud Storage como repositorio de documentos PDF.
- La integración con el modelo generativo Gemini para la generación de respuestas en lenguaje natural.

Este diseño permitió definir una arquitectura modular, escalable y coherente con los requerimientos institucionales.

b) Creación y configuración del entorno cloud

En esta etapa se abordó la configuración del entorno cloud sobre Google Cloud Platform (GCP), definiendo la infraestructura base que soporta la solución.

Las principales tareas realizadas incluyeron:

- Configuración del servicio Google Cloud Storage para almacenamiento de documentos PDF.
- Configuración de Cloud SQL (PostgreSQL) como base de datos relacional.
- Definición de políticas de acceso y permisos mediante Identity and Access Management (IAM).

Esta etapa permitió disponer de un entorno operativo seguro y funcional sobre el cual integrar los flujos de la plataforma.

c) Implementación del flujo de carga y almacenamiento

Se implementó el flujo de carga de información y documentos coordinado por el backend desarrollado en Django + Django REST Framework, incluyendo:

- Recepción de cargas desde el frontend Angular.
- Registro de metadatos en la base de datos relacional (Cloud SQL) mediante Django ORM.
- Operaciones de carga y descarga de archivos en Google Cloud Storage.

Esta etapa validó la correcta integración entre la aplicación y los servicios cloud, asegurando trazabilidad y consistencia de la información institucional.

d) Integración del asistente conversacional con Gemini

Se integró un asistente conversacional que permite realizar consultas en lenguaje natural sobre información institucional estructurada.

El backend actúa como intermediario entre los datos institucionales y el modelo generativo, realizando:

1. Interpretación de la consulta del usuario.
2. Recuperación de información estructurada desde la base de datos relacional.
3. Construcción de un contexto textual controlado.
4. Envío del contexto al modelo Gemini para generar respuestas en lenguaje natural.

En esta implementación, el modelo generativo no accede directamente a los documentos PDF, ni realiza recuperación semántica documental, validando la factibilidad técnica de incorporar un asistente inteligente sin exponer directamente los repositorios institucionales.

Síntesis del aporte individual

El aporte individual del trabajo se centra en el diseño e implementación de la infraestructura cloud que sustenta la gestión documental y la consulta inteligente de información institucional.

La tesina demuestra cómo la combinación de:

- Servicios cloud administrados,
- Una arquitectura modular,
- Y modelos de lenguaje generativo,

permite modernizar los procesos administrativos y académicos de la Academia Nacional de Bomberos de Chile, sentando las bases técnicas para una futura evolución hacia soluciones de búsqueda semántica avanzada sobre documentos institucionales.

3.6 Consideraciones de seguridad

La seguridad y privacidad de la información constituye un aspecto transversal dentro del diseño de la solución propuesta.
En el marco teórico de este trabajo se revisaron los principales principios, estándares y buenas prácticas de seguridad aplicables a entornos cloud y gestión de información institucional, los cuales sirven como base conceptual para la definición de los mecanismos de protección incorporados en la arquitectura diseñada.

Desde el alcance de esta etapa de implementación, la seguridad se aborda como un requisito técnico de la infraestructura cloud, asegurando que el flujo de carga, almacenamiento y consulta de información se ejecute bajo principios de confidencialidad, integridad y disponibilidad.

Seguridad en el entorno cloud

La arquitectura se implementa sobre Google Cloud Platform (GCP), aprovechando las capacidades de seguridad provistas por los servicios administrados utilizados.
Los datos son protegidos tanto en tránsito como en reposo, mediante mecanismos de cifrado gestionados por la plataforma, lo que reduce el riesgo de accesos no autorizados o interceptaciones externas.

El control de acceso a los recursos se gestiona a través de Identity and Access Management (IAM), permitiendo asignar permisos específicos a cada componente del sistema.
Este enfoque garantiza que solo los servicios y procesos autorizados puedan interactuar con la base de datos, el almacenamiento de documentos y los servicios de inteligencia artificial, fortaleciendo la trazabilidad y el control operativo del sistema.

Seguridad en la integración con modelos de lenguaje

El asistente inteligente basado en Gemini se integra de forma controlada a través del backend, el cual actúa como intermediario entre los datos institucionales y el modelo generativo.
En la implementación actual, el modelo no accede directamente a los documentos ni a los repositorios de almacenamiento, sino que recibe únicamente un contexto textual construido por el backend, limitando la exposición de información sensible.

En la arquitectura propuesta como evolución, el uso de técnicas de Retrieval-Augmented Generation (RAG) considera accesos temporales y controlados al contenido procesado, manteniendo la separación entre los datos originales y los mecanismos de generación de respuestas.

Cumplimiento y buenas prácticas

El diseño de la arquitectura cloud considera buenas prácticas ampliamente aceptadas en la gestión de información institucional, tales como la mínima exposición de datos, la

separación de responsabilidades entre componentes y el uso de servicios administrados para reducir riesgos operativos.

Si bien el análisis detallado de estándares y normativas de seguridad se aborda a nivel conceptual en el marco teórico, la solución propuesta se alinea con dichos principios mediante la correcta configuración de los servicios cloud y el control de accesos, asegurando un entorno de procesamiento seguro y confiable.

De esta manera, la solución propuesta materializa en la arquitectura cloud los principios de seguridad revisados en el marco teórico, mediante el uso de cifrado nativo de los servicios administrados, control de accesos mediante IAM y separación de responsabilidades entre componentes.

Si bien no se realiza en esta sección un análisis detallado de estándares o normativas específicas, dichos aspectos fueron abordados conceptualmente en el marco teórico, mientras que en esta sección se presenta su aplicación práctica dentro de la arquitectura implementada, validando la coherencia entre el diseño técnico y los principios de seguridad definidos.

4 Validación de la solución

4.1 Estrategia de validación

La validación de la solución se realizó mediante pruebas funcionales sobre la aplicación web operativa desarrollada como parte del proyecto. Estas pruebas tuvieron como objetivo comprobar el correcto funcionamiento de los módulos implementados en la plataforma y la integración del asistente conversacional basado en modelos generativos.

De manera complementaria, se realizaron pruebas técnicas controladas para explorar la factibilidad de incorporar en el futuro mecanismos de recuperación semántica documental (enfoque RAG). Dado que este componente no forma parte de la implementación validada en esta etapa, dichas pruebas se ejecutaron de forma aislada mediante el consumo directo de endpoints REST utilizando la herramienta Postman, sin integración con la aplicación web productiva.

En consecuencia, la validación presentada en este capítulo se distingue en:

• Validación funcional: enfocada en la aplicación web y el asistente conversacional implementado, comprobando los flujos de carga, almacenamiento, consulta y visualización de información institucional.

• Validación técnica exploratoria: enfocada en verificar de manera aislada la viabilidad de integración futura de servicios de recuperación semántica y modelos generativos sobre documentos institucionales.

En este contexto, la validación se plantea como funcional, en cuanto al correcto comportamiento de la aplicación web y el asistente conversacional, y técnica, respecto a la factibilidad de integrar mecanismos de búsqueda semántica avanzada sobre documentos institucionales. No se consideran en esta etapa evaluaciones con usuarios

finales ni mediciones de rendimiento a gran escala, las cuales se proponen como trabajo futuro.

4.2 Validación funcional de la aplicación web

La validación funcional de la solución se realizó mediante el uso directo de la aplicación web desarrollada, verificando el correcto funcionamiento de los principales módulos implementados en la plataforma *Bomberos Academia*. Estas pruebas permitieron comprobar que el sistema cumple con los requerimientos definidos para la gestión de información institucional y la interacción con el asistente inteligente.

En particular, se validaron los siguientes aspectos funcionales:

- **Gestión de perfiles de bomberos**: Se comprobó la correcta creación, visualización y actualización de los perfiles institucionales, asegurando la persistencia de los datos en la base de datos relacional gestionada mediante Django ORM. Esta validación confirma la correcta integración entre el backend y el sistema de persistencia basado en PostgreSQL (Cloud SQL).

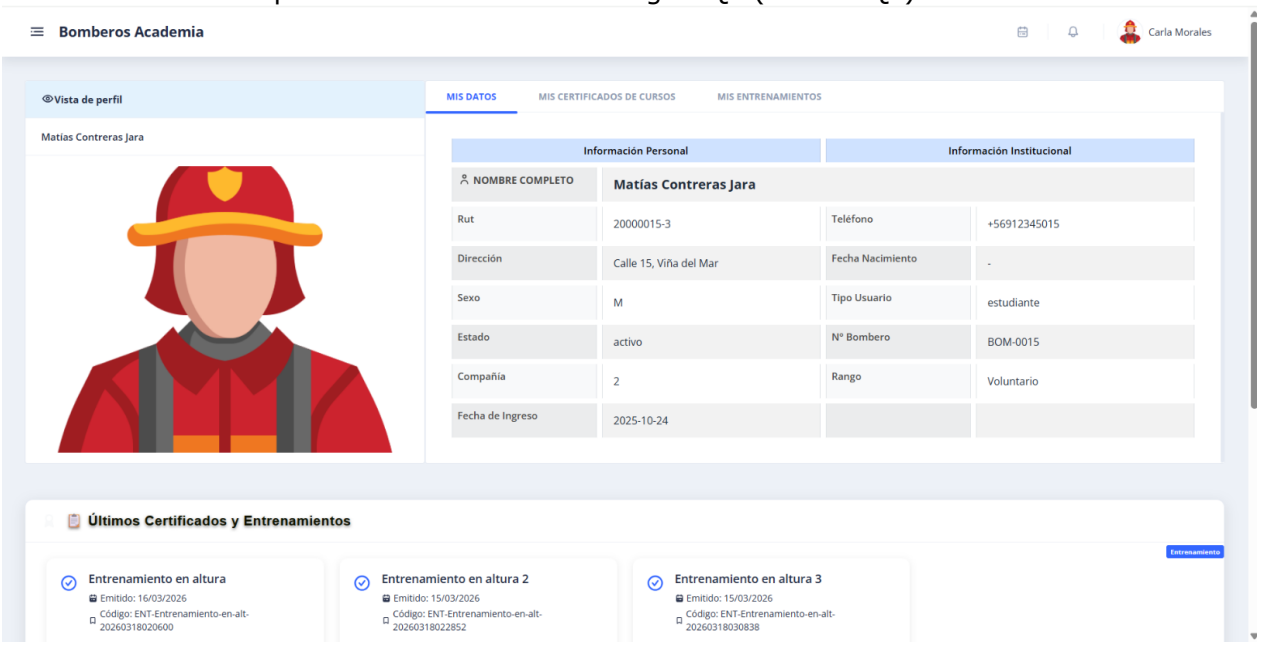


Figura 9. Vista de perfil de bombero en la plataforma.

Fuente: Elaboración propia

Vista del perfil de un bombero en la plataforma que muestra la información personal e institucional almacenada en la base de datos, junto con los certificados y entrenamientos asociados, la creación y asociación de dichos certificados será evidenciada en las siguientes figuras.

- **Gestión académica**: Se verificó el funcionamiento de los módulos asociados a cursos, módulos y certificaciones, confirmando la correcta asociación de estos elementos a los perfiles correspondientes y su posterior consulta desde la interfaz web.

- **Gestión documental**: Se validó la carga, almacenamiento y recuperación de documentos institucionales (por ejemplo, certificados en formato PDF), comprobando su correcta asociación a los registros académicos y su disponibilidad para visualización o descarga desde la plataforma.
  Los archivos fueron almacenados en Google Cloud Storage, mientras que los metadatos y relaciones se mantuvieron en la base de datos relacional.
  Para respaldar este flujo, se presentan las siguientes evidencias:

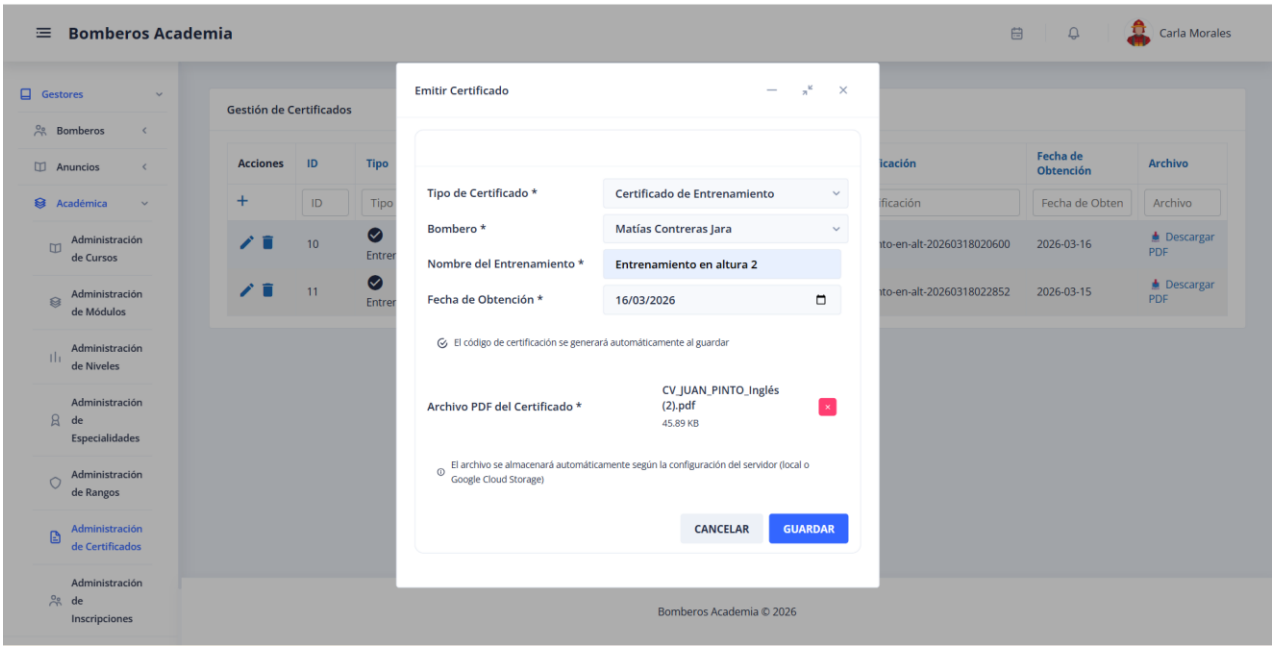


Figura 10. Carga de certificado nuevo.

Fuente: Elaboración propia

Formulario de carga de certificados desde la interfaz web, donde el usuario ingresa los datos asociados y selecciona el archivo PDF a subir.

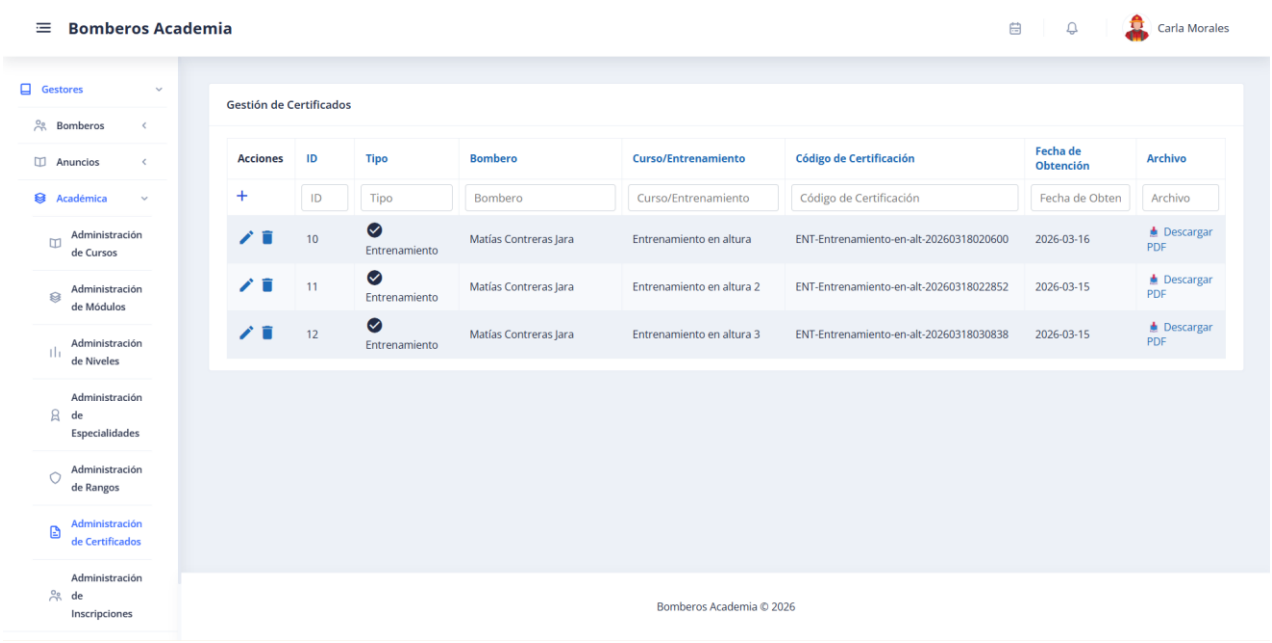


Figura 11. Vista de página web posterior a la carga.

Fuente: Elaboración propia

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

INFO 2026-03-17 23:22:20,568 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:23:20,570 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:23:49,684 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:23:57,582 models AFC is enabled with max remote calls: 10.
INFO 2026-03-17 23:23:58,588 _client HTTP Request: POST https://generativelanguage.googleapis.com/v1beta/models/gemini-2.5-flash:generateConten
t "HTTP/1.1 200 OK"
INFO 2026-03-17 23:23:58,590 basehttp "POST /api/gemini/chat/ HTTP/1.1" 200 427
INFO 2026-03-17 23:24:03,548 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
WARNING 2026-03-17 23:24:32,717 log Forbidden: /api/eventos/notificaciones/mis-notificaciones/
WARNING 2026-03-17 23:24:32,717 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 403 172
INFO 2026-03-17 23:24:32,742 basehttp "POST /api/token/refresh/ HTTP/1.1" 200 358
INFO 2026-03-17 23:24:32,768 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:25:03,544 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:25:32,725 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:26:03,558 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:26:33,547 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:27:20,580 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:27:54,925 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:28:02,717 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:28:05,264 basehttp "GET /api/user/perfil/navbar/ HTTP/1.1" 200 35
INFO 2026-03-17 23:28:05,270 basehttp "GET /api/cursos/certificados/ HTTP/1.1" 200 582
INFO 2026-03-17 23:28:05,273 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:28:05,289 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
INFO 2026-03-17 23:28:16,068 basehttp "GET /api/cursos/ HTTP/1.1" 200 1876
INFO 2026-03-17 23:28:16,157 basehttp "GET /api/user/bomberos/estudiantes/ HTTP/1.1" 200 17633
INFO 2026-03-17 23:28:35,277 basehttp "GET /api/eventos/notificaciones/mis-notificaciones/ HTTP/1.1" 200 45
📄 Procesando archivo: CV_JUAN_PINTO_Inglés (2).pdf
🔧 GCS configurado: True
🪣 Bucket: bomberos-academia-media
🔑 USE_GCS_STORAGE: True
📤 Subiendo a: media/certificados/ENT-Entrenamiento-en-alt-20260318022852_CV_JUAN_PINTO_Inglés (2).pdf
📤 Iniciando subida a GCS...
   - Bucket: bomberos-academia-media
   - Destino: media/certificados/ENT-Entrenamiento-en-alt-20260318022852_CV_JUAN_PINTO_Inglés (2).pdf
   - Archivo: CV_JUAN_PINTO_Inglés (2).pdf
   - Content-Type: application/pdf
   - Subiendo archivo...
```

Figura 12. Proceso de carga de documento desde el backend hacia Google Cloud Storage.

Fuente: Elaboración propia

Registro de ejecución del backend durante la carga de un documento PDF, evidenciando la configuración del servicio, el bucket de destino y la ruta de almacenamiento en Google Cloud Storage.

Asimismo, los registros evidencian llamadas HTTPS al servicio generativo Gemini, validando la comunicación segura en tránsito con servicios externos.

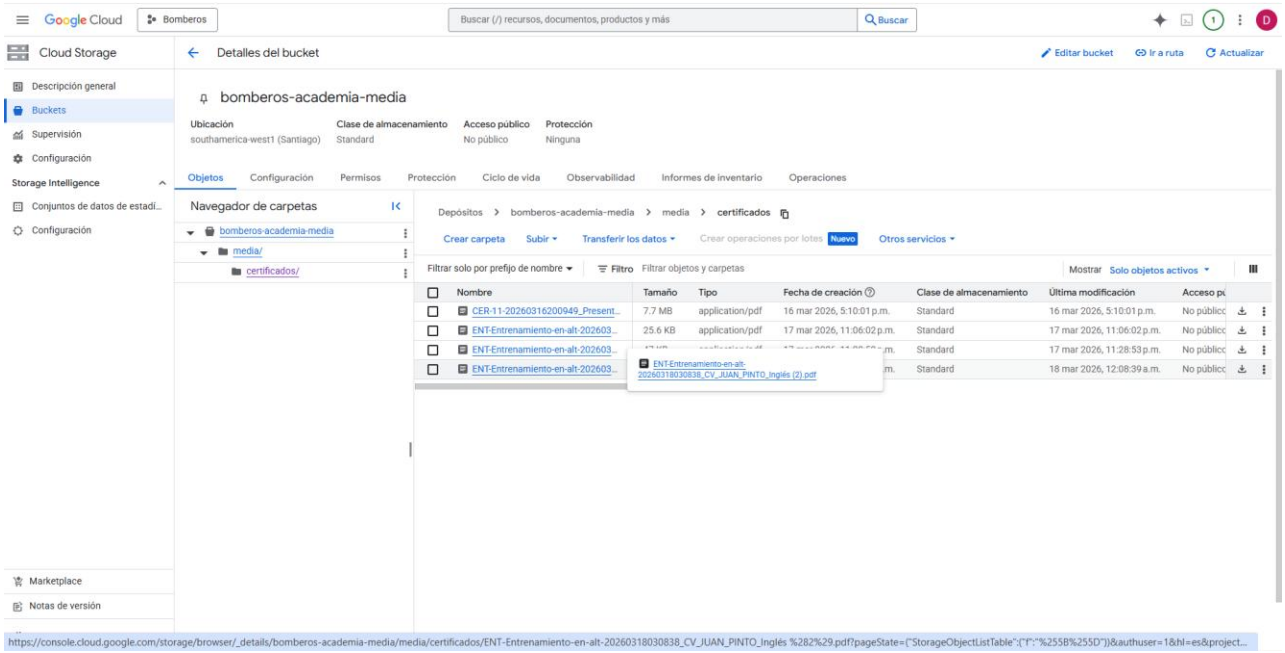


Figura 13. Vista del documento en bucket de Cloud Storage posterior a la carga.

Fuente: Elaboración propia

Vista del bucket de Google Cloud Storage que muestra el documento almacenado tras su carga desde la aplicación web, incluyendo su nombre, tipo y fecha de creación.

La evidencia presentada permite validar no solo el funcionamiento funcional del sistema, sino también la correcta integración entre el frontend, el backend y el servicio de almacenamiento cloud, evidenciando un flujo completo desde la carga del documento por parte del usuario hasta su almacenamiento efectivo en la infraestructura de Google Cloud.

- **Interacción con el asistente conversacional**: Se probó el módulo de chatbot integrado en la interfaz web, el cual permite a los usuarios realizar consultas en lenguaje natural.

  En estas pruebas, el backend procesó las solicitudes, obtuvo la información estructurada desde la base de datos y envió el contexto correspondiente al modelo generativo Gemini, obteniendo respuestas coherentes y alineadas con la información institucional disponible.

  Para validar este flujo, se presentan las siguientes evidencias:

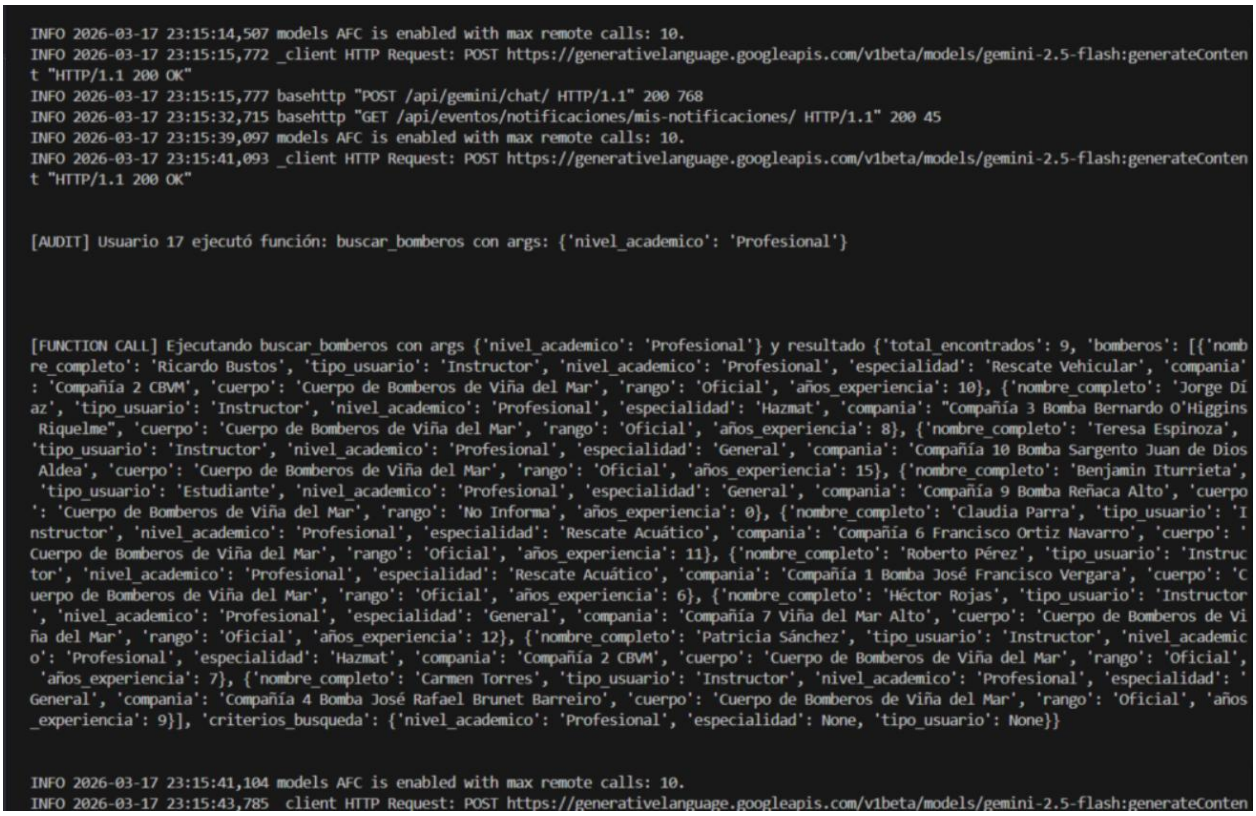


Figura 14. Ejecución de llamada del backend al servicio Gemini.

Fuente: Elaboración propia

Registro de ejecución del backend que evidencia una solicitud HTTP POST al servicio Gemini, con respuesta exitosa (HTTP 200), validando la conectividad entre la aplicación y el modelo generativo.

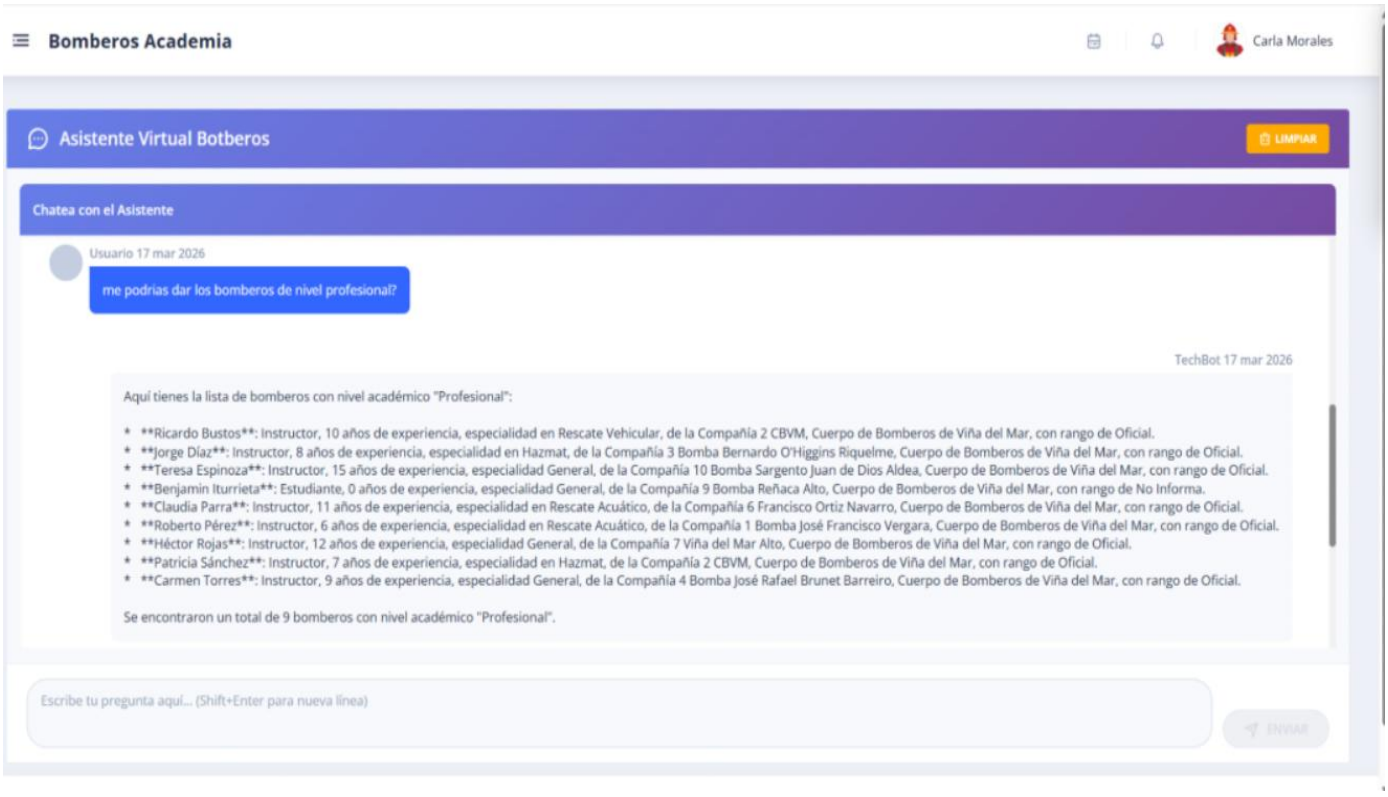


Figura 15. Ejemplo de conversación y consulta al chatbot.

Fuente: Elaboración propia

Ejemplo de consulta realizada por un usuario en la interfaz web, junto con la respuesta generada por el asistente conversacional a partir de información estructurada del sistema.

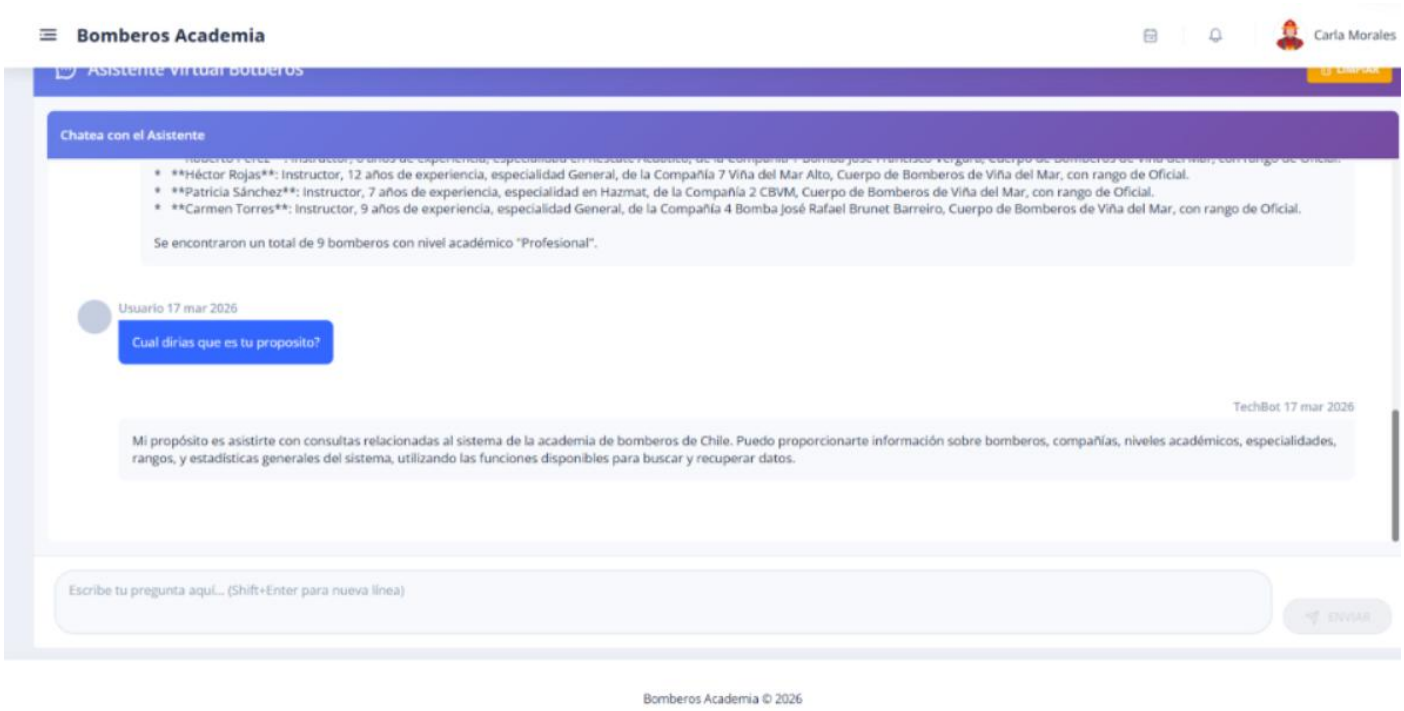


Figura 16. Continuación de ejemplo de conversación.

Fuente: Elaboración propia

Estas evidencias permiten validar la correcta integración entre el backend de la aplicación y los servicios de inteligencia artificial en la nube, demostrando la capacidad del sistema para procesar consultas en lenguaje natural y generar respuestas basadas en datos institucionales.

- **Validación de API y flujo end-to-end**: Se validó el correcto funcionamiento de la API REST mediante la observación de respuestas HTTP exitosas (códigos 200 y 201), evidenciando la correcta comunicación entre el frontend y el backend.

  Asimismo, se verificó el flujo completo del sistema (end-to-end), desde la interacción del usuario en la interfaz web, pasando por el procesamiento en el backend, la persistencia en la base de datos o almacenamiento en servicios cloud, hasta la visualización de resultados en la interfaz.

- **Conclusión de la validación funcional**: Los resultados de estas pruebas evidencian que la aplicación web es funcional y operativa, permitiendo ejecutar de forma integrada los principales flujos del sistema, desde la gestión de información académica y documental hasta la consulta mediante un asistente inteligente.

Esta validación confirma:

- la correcta implementación de la arquitectura cliente–servidor
- la integración efectiva con servicios cloud (Cloud SQL, Cloud Storage y Gemini)
- la viabilidad del uso de modelos generativos como apoyo a la consulta de información institucional

4.3 Validación técnica exploratoria del flujo RAG

Con el fin de evaluar la factibilidad técnica de incorporar en el futuro mecanismos de recuperación semántica documental, se realizaron pruebas controladas sobre servicios de Vertex AI, consumiendo directamente endpoints desde la herramienta Postman.

Estas pruebas consistieron en enviar consultas en lenguaje natural a un endpoint de prueba que integra un flujo básico de recuperación de fragmentos relevantes desde documentos previamente indexados y generación de respuestas mediante el modelo generativo Gemini.

La Figura 17 muestra un ejemplo de respuesta obtenida a partir de una consulta de prueba, evidenciando la correcta interacción entre el backend de prueba y los servicios de Vertex AI.

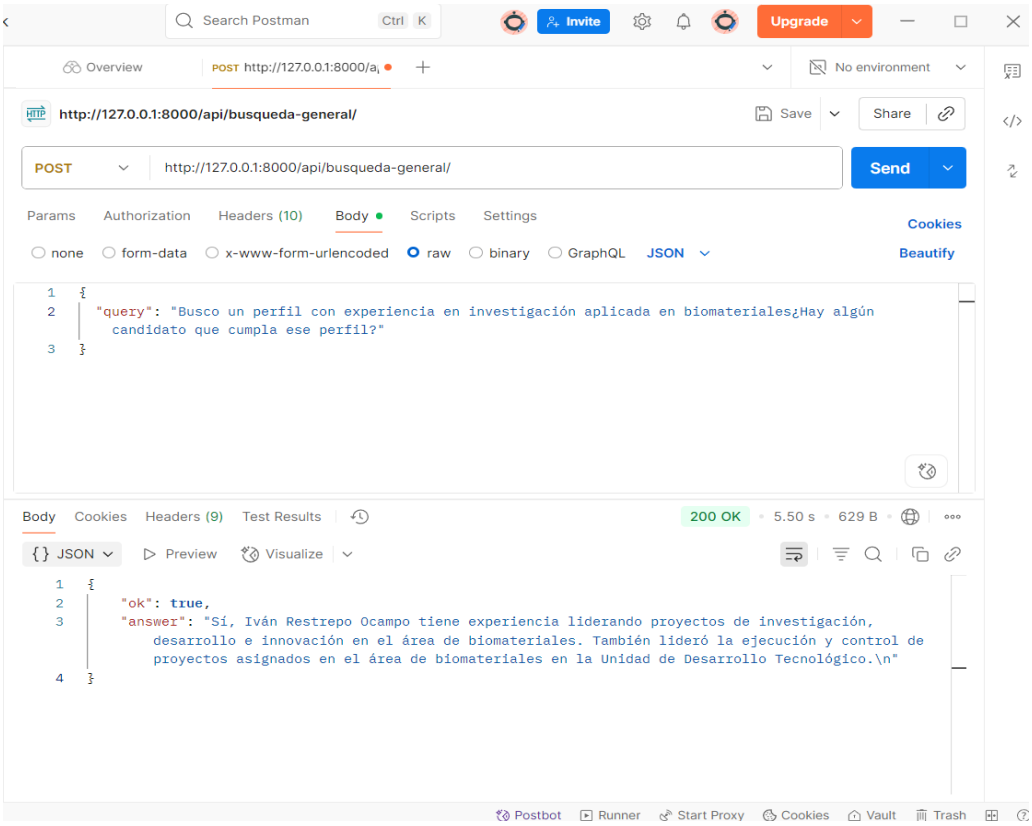


Figura 17. Resultado de Prueba técnica exploratoria del endpoint de recuperación semántica utilizando Vertex AI – Gemini.

Fuente: Elaboración propia

Es importante destacar que este flujo no se encuentra integrado actualmente en la aplicación web productiva, sino que corresponde a una validación técnica aislada que respalda la factibilidad de evolución futura hacia una arquitectura RAG completa.

4.4 Análisis y discusión de resultados

Los resultados de la validación funcional confirman el correcto funcionamiento de la plataforma web implementada y del asistente conversacional basado en Gemini. Por su parte, las pruebas técnicas exploratorias respaldan la factibilidad de incorporar en el futuro mecanismos de recuperación semántica documental mediante una arquitectura RAG, sin que estos formen parte de la implementación validada en la presente etapa.

Los resultados obtenidos durante la validación del prototipo evidencian que la solución propuesta cumple con los objetivos funcionales definidos para esta etapa del trabajo de título. La plataforma web desarrollada permite gestionar correctamente la información institucional estructurada —perfiles de bomberos, cursos y certificaciones— a través de una arquitectura cliente–servidor basada en Django y una base de datos relacional, asegurando coherencia, trazabilidad y disponibilidad de los datos.

Por otra parte, en el contexto de las pruebas técnicas exploratorias, la utilización de Vertex AI con Gemini permitió generar respuestas coherentes y contextualizadas a partir de fragmentos de información recuperados desde documentos previamente indexados en el entorno de prueba, demostrando la factibilidad técnica de esta línea de evolución.

Si bien en la implementación actual el modelo generativo no accede directamente a los documentos almacenados en Cloud Storage, los resultados obtenidos validan el flujo de construcción de contexto desde datos estructurados, sentando una base sólida para la futura incorporación de recuperación semántica directa sobre documentos (enfoque RAG completo). Esta decisión de diseño permite desacoplar la lógica del sistema, controlar costos y reducir la complejidad técnica en una etapa temprana del proyecto.

Desde una perspectiva técnica, los resultados demuestran que la arquitectura cloud propuesta es escalable, modular y consistente con buenas prácticas de desarrollo en la nube. No obstante, se identifican como limitaciones actuales la ausencia de métricas de rendimiento a gran escala y la falta de evaluación con usuarios finales, aspectos que se proponen abordar en trabajos futuros.

En síntesis, la validación realizada confirma que la solución es técnicamente viable, funcionalmente correcta y adecuada como base para la evolución hacia un sistema de gestión documental con capacidades avanzadas de búsqueda semántica y recuperación aumentada de información.

5 Conclusiones y trabajos futuros

5.1 Conclusiones

El presente trabajo de título abordó el diseño e implementación de una infraestructura cloud orientada a la gestión documental y consulta inteligente de información institucional para la Academia Nacional de Bomberos de Chile. A partir del análisis del contexto y la problemática existente, se propuso una solución tecnológica basada en arquitecturas cloud-native, que permite centralizar la información académica y administrativa, mejorar su trazabilidad y habilitar mecanismos de consulta en lenguaje natural.

Los resultados obtenidos demuestran que la plataforma desarrollada cumple con los objetivos definidos para esta etapa del proyecto. El sistema permite gestionar de manera efectiva perfiles, cursos, certificaciones y documentos institucionales mediante una aplicación web funcional, respaldada por un backend en Django y una base de datos relacional. Esta arquitectura asegura consistencia de los datos, control de accesos y una correcta separación de responsabilidades entre los distintos componentes del sistema.

Asimismo, se validó la integración de un asistente inteligente basado en Gemini para responder consultas en lenguaje natural a partir de información estructurada proporcionada por el backend. De manera complementaria, las pruebas técnicas exploratorias realizadas sobre Vertex AI respaldan la factibilidad de una evolución futura hacia mecanismos avanzados de recuperación semántica documental.

Desde el punto de vista de ingeniería, el trabajo permitió aplicar y consolidar conocimientos relacionados con arquitecturas en la nube, diseño de APIs, gestión de datos y servicios administrados de inteligencia artificial. Además, se definió una arquitectura modular y escalable que permite evolucionar el sistema sin afectar su operación actual.

Adicionalmente, el análisis cuantitativo de costos realizado en el marco conceptual permitió estimar el gasto asociado a la infraestructura cloud base necesaria para operar la plataforma. Los resultados evidenciaron que la utilización de servicios administrados en Google Cloud Platform permite implementar la solución con costos operativos acotados y predecibles, especialmente en escenarios institucionales con alta demanda de almacenamiento y transferencia de documentos. Este análisis respalda la viabilidad económica de la implementación propuesta y confirma que el despliegue de la plataforma en un entorno cloud resulta técnica y financieramente factible para la Academia Nacional de Bomberos de Chile.

En conclusión, el trabajo desarrollado constituye una base tecnológica sólida para la modernización de los procesos administrativos y académicos de la Academia Nacional de Bomberos de Chile, demostrando que el uso de infraestructuras cloud y modelos de inteligencia artificial es una alternativa viable y pertinente en contextos institucionales reales.

5.2 Limitaciones del trabajo

Si bien la solución cumple con los objetivos propuestos, existen algunas limitaciones que deben ser consideradas:

- La validación del sistema se realizó principalmente a nivel funcional y técnico, sin incluir aún evaluaciones con usuarios finales en un entorno productivo institucional.
- El módulo de búsqueda inteligente actualmente construye su contexto únicamente a partir de información estructurada, sin incorporar recuperación semántica directa sobre documentos PDF.
- No se realizaron pruebas de carga ni mediciones de desempeño a gran escala, las cuales serían necesarias para un despliegue institucional completo.

5.3 Trabajos futuros

A partir de los resultados obtenidos, se identifican diversas líneas de trabajo futuro:

- Implementar un flujo completo de recuperación semántica sobre documentos institucionales mediante un enfoque Retrieval-Augmented Generation (RAG), integrando directamente Cloud Storage con servicios de indexación y búsqueda vectorial.
- Incorporar métricas de rendimiento, monitoreo y logging para evaluar el comportamiento del sistema bajo carga real.
- Realizar pruebas piloto con usuarios institucionales para evaluar la usabilidad, utilidad y aceptación del sistema.
- Extender la plataforma con módulos adicionales de analítica, reportes avanzados y control automatizado de certificaciones.
- Evaluar e integrar explícitamente estándares de seguridad y normativas de gestión documental aplicables al contexto institucional.

Referencias:

[1] ISO 15489-1:2016. *Information and documentation — Records management — Part 1: Concepts and principles*. International Organization for Standardization, Geneva, 2016.
Available at: https://www.iso.org/standard/62542.html

[2] Al-Hussein, M., Al-Khateeb, H., Qatawneh, M.: *AI-based Integrated Approach for the Development of Intelligent Document Management Systems*. Procedia Computer Science, Vol. 219, 2023, pp. 485–492.
Available at: https://www.sciencedirect.com/science/article/pii/S1877050923021324

[3] IBM Corporation: *Intelligent Document Processing: AI-powered automation for document workflows*. IBM Think, 2020.
Available at: https://www.ibm.com/think/topics/intelligent-document-processing

[4] Mell, P., Grance, T.: The NIST Definition of Cloud Computing.
NIST Special Publication 800-145, National Institute of Standards and Technology, 2011.
Available at: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[5] Armbrust, M., et al.: A View of Cloud Computing.
Communications of the ACM, Vol. 53, No. 4, 2010, pp. 50–58.
Available at: https://dl.acm.org/doi/10.1145/1721654.1721672

[6] Ghemawat, S., Gobioff, H., Leung, S.-T.: The Google File System.
ACM Symposium on Operating Systems Principles (SOSP), 2003.
Available at: https://research.google/pubs/pub51/

[7] Manning, C. D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, 2008.
Available at: https://nlp.stanford.edu/IR-book/

[8] Lewis, P., Perez, E., Piktus, A., et al.: *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Advances in Neural Information Processing Systems (NeurIPS), 2020.
Available at: https://arxiv.org/abs/2005.11401

[9] Gao, Y., et al.: *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv preprint arXiv:2312.10997, 2023.
Available at: https://arxiv.org/abs/2312.10997

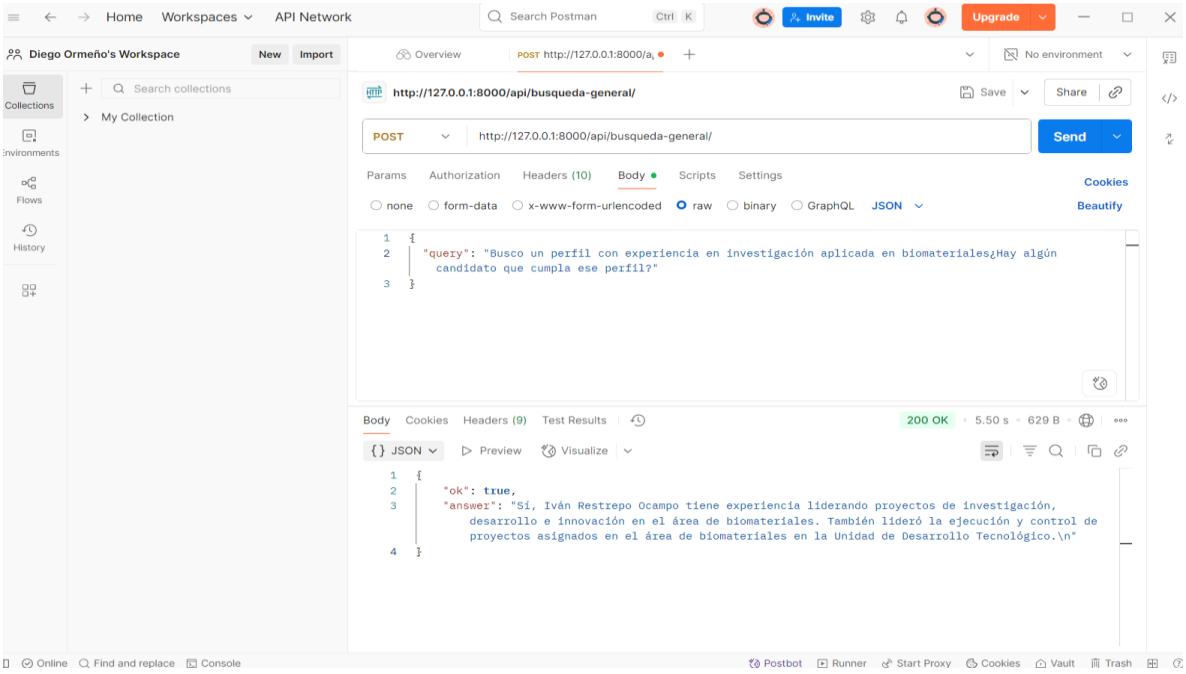
[10] Amazon Web Services: AWS Cloud Computing Overview.
AWS Documentation, 2023.
Available at: https://aws.amazon.com/what-is-cloud-computing/

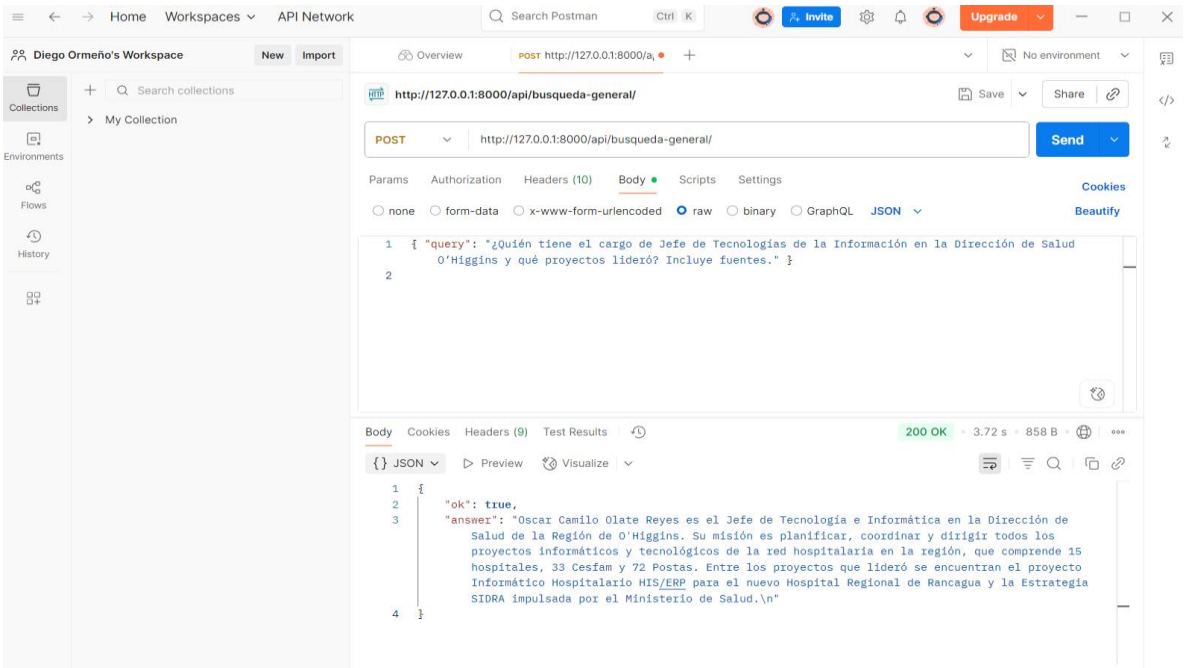
[11] Microsoft Azure: *Azure AI Services Documentation*. Microsoft, 2023.
Available at: https://learn.microsoft.com/azure/ai-services/

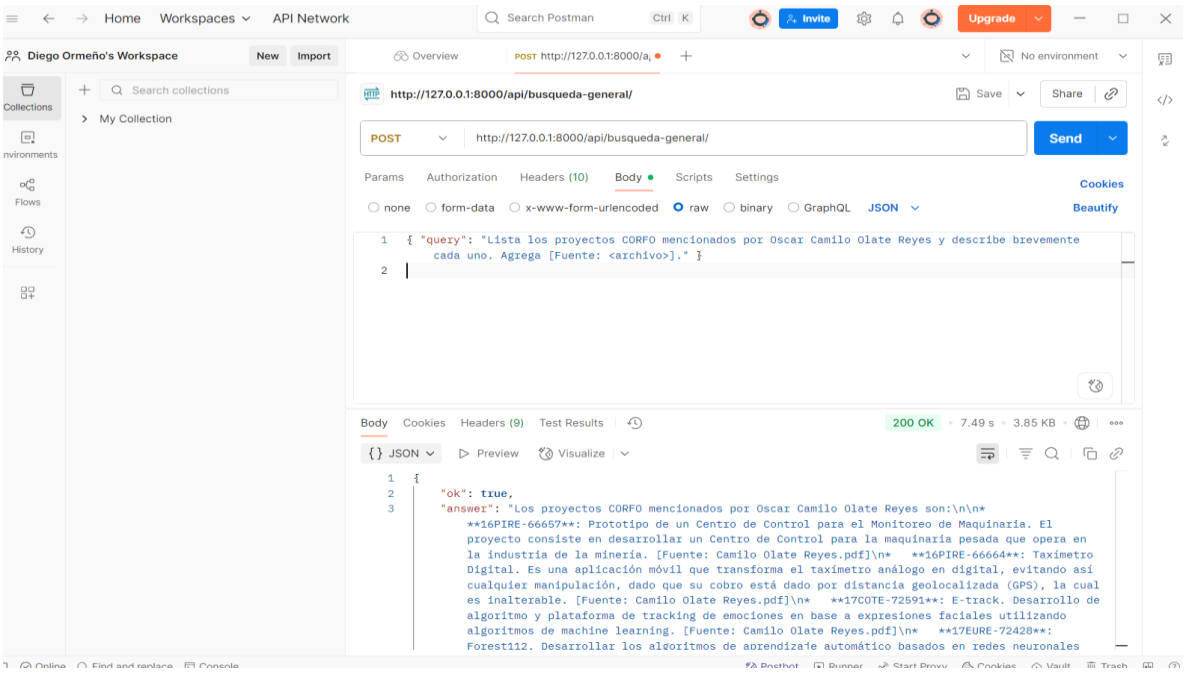
[12] Google Cloud: *Vertex AI and Document AI Overview*. Google Cloud Documentation, 2024.
Available at: https://cloud.google.com/vertex-ai

Anexos:

Anexo A: Pruebas de recuperación semántica utilizando Vertex AI – Gemini:






Home
Workspaces
API Network
Search Postman
Invite
Upgrade
Diego Ormeño's Workspace
New
Import
Collections
Flows
History
Search collections
My Collection
Overview
No environment
http://127.0.0.1:8000/api/busqueda-general/
Save
Share
POST
http://127.0.0.1:8000/api/busqueda-general/
Send
Params
Authorization
Headers (10)
Body
Scripts
Settings
Cookies
none
form-data
x-www-form-urlencoded
raw
binary
GraphQL
JSON
Beautify
{ "query": "Lista los proyectos CORFO mencionados por Oscar Camilo Olate Reyes y describe brevemente cada uno. Agrega [Fuente: <archivo>]." }
Body
Cookies
Headers (9)
Test Results
200 OK
7.49 s
3.85 KB
JSON
Preview
Visualize
{
"ok": true,
"answer": "Los proyectos CORFO mencionados por Oscar Camilo Olate Reyes son:\n\n* **16PIRE-66657**: Prototipo de un Centro de Control para el Monitoreo de Maquinaria. El proyecto consiste en desarrollar un Centro de Control para la maquinaria pesada que opera en la industria de la minería. [Fuente: Camilo Olate Reyes.pdf]\n* **16PIRE-66664**: Taxímetro Digital. Es una aplicación móvil que transforma el taxímetro análogo en digital, evitando así cualquier manipulación, dado que su cobro está dado por distancia geolocalizada (GPS), la cual es inalterable. [Fuente: Camilo Olate Reyes.pdf]\n* **17COTE-72591**: E-track. Desarrollo de algoritmo y plataforma de tracking de emociones en base a expresiones faciales utilizando algoritmos de machine learning. [Fuente: Camilo Olate Reyes.pdf]\n* **17EURE-72428**: Forest112. Desarrollar los algoritmos de aprendizaje automático basados en redes neuronales