Análisis automático de buenas prácticas de arquitectura en implementaciones en Google Cloud Platform

Roger Steven Ramírez Espejo

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERÍA

MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

BOGOTÁ

20 de noviembre de 2024

Análisis automático de buenas prácticas de arquitectura en implementaciones en Google Cloud Platform

Roger Steven Ramírez Espejo

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Directora:

Ing. Mariela Josefina Curiel Huérfano, Ph.D.

Comité de Evaluación del Trabajo de Grado:

Andrea del Pilar Rueda Olarte, Ph.D.

Jaime Andrés Pavlich Mariscal, Ph.D.

Página web del trabajo de Grado

`https://github.com/roger8849/gcp_infra_best_practices_analyzer`

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERÍA

MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

BOGOTÁ

20 de noviembre de 2024

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERÍA

MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Luis Fernando Múnera Congote, S.J.

Decano Facultad de Ingeniería

Ing. Lope Hugo Barrero Solano, Sc.D.

Directora Maestría en Ingeniería de Sistemas y Computación

Ing. Mariela Josefina Curiel Huérfano, Ph.D.

Director Departamento de Ingeniería de Sistemas

Ing. César Julio Bustacara Medina, Ph.D.

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

Dedicatoria.

”No puedo entender por qué la gente tiene miedo de las nuevas ideas. Yo tengo miedo de las viejas.”

Alan Turing

Agradecimientos

Quiero aprovechar este espacio para expresar mi más sincero agradecimiento a mi directora de trabajo de grado, Mariela Curiel Huérfano, por su invaluable orientación y apoyo a lo largo de todo el proceso de escritura y desarrollo de este proyecto. Su dedicación, paciencia y comprensión han sido fundamentales para el logro de este trabajo académico. Durante el tiempo que hemos trabajado juntos, la profesora Mariela ha demostrado un profundo conocimiento en el campo de estudio, brindándome una guía experta y ofreciéndome valiosas perspectivas que han enriquecido mi trabajo. Sus comentarios y sugerencias han sido fundamentales para mejorar la calidad de este trabajo final, así como para impulsar mi crecimiento académico y profesional.

Deseo expresar mi profundo agradecimiento a mi familia: mis padres, José y Cecilia y mis Hermanos Jenny, Carol y Cesar, quienes desde su juventud enfrentaron grandes desafíos para formar a nuestra familia en un país que tanto en la actualidad y en aquel entonces, presenta limitaciones significativas en cuanto a oportunidades de desarrollo y educación. Su perseverancia, dedicación y valentía allanaron el camino para que pudiera tener una vida más próspera y llena de oportunidades. Reconozco que mis logros alcanzados en el ámbito educativo no son solo míos, sino que son también el resultado del sacrificio y el apoyo incondicional de mis padres. La determinación y el esfuerzo incansable de mis padres me inspiraron a perseguir mis metas académicas y a superar cualquier obstáculo que se presentara en el camino.

También deseo expresar mi más sincero agradecimiento a mi hijo Matías y a mi mascota Peggy, por ser mi constante fuente de inspiración y apoyo incondicional. Su amor, alegría y presencia en mi vida ha sido fundamental para mantenerme motivado y superar los desafíos que he enfrentado. En los momentos más difíciles, su presencia ha sido un bálsamo que me ha dado fuerzas para seguir adelante.

Finalmente, quiero dar las gracias a todos los profesores, colegas, y alumnos que han participado en mis 18 años de formación académica y profesional. Sin su ayuda, no habría podido llegar hasta donde estoy hoy.

Índice general

Abstract

Automatic Analysis of Best Architecture Practices in Implementations on Google Cloud

This work presents the development and evaluation of an application leveraging natural language models to analyze network configurations in Google Cloud Platform (GCP) and recommend improvements based on best practices. The application collects data from GCP resources, such as VPCs, VPNs, firewalls, and subnets, and processes them using language models to generate a detailed report. The report highlights specific and general recommendations derived from public GCP documentation to enhance network security, efficiency, and maintainability. A simulated GCP organization (*roma.joonix.net*) was created to evaluate the tool, reflecting a realistic cloud topology with hub-and-spoke network models and multi-project configurations. Results indicate that the application successfully identifies potential improvements, such as simplifying redundant network configurations and addressing security vulnerabilities like permissive firewall rules. While some recommendations were specific to the provided configurations, others were generic, indicating the need for more context in some scenarios. The study concludes that language models are effective for generating actionable recommendations in cloud environments. However, future iterations should integrate a hybrid approach combining deterministic methods with language-based analysis to enhance precision and contextual relevance. Expanding the tool's scope to include other GCP services is also recommended.

Keywords: Google Cloud Platform, network configurations, VPC, VPN, best practices, language models, cloud security, network optimization, hybrid analysis.

Resumen

Análisis automático de buenas prácticas de arquitectura en implementaciones en Google Cloud Platform

Este trabajo presenta el desarrollo y la evaluación de una aplicación que utiliza modelos de lenguaje natural para analizar configuraciones de red en Google Cloud Platform (GCP) y recomendar mejoras basadas en buenas prácticas. La aplicación recopila datos de recursos de GCP, como VPC, VPN, firewalls y subredes, y los procesa utilizando modelos de lenguaje para generar un informe detallado. El informe resalta recomendaciones específicas y generales derivadas de la documentación pública de GCP para mejorar la seguridad, la eficiencia y el mantenimiento de la red. Para evaluar la herramienta, se creó una organización simulada de GCP (*roma.joonix.net*) que refleja una topología de nube realista con modelos de red hub-and-spoke y configuraciones de múltiples proyectos. Los resultados indican que la aplicación identifica con éxito posibles mejoras, como la simplificación de configuraciones de red redundantes y la resolución de vulnerabilidades de seguridad, como reglas de firewall demasiado permisivas. Si bien algunas recomendaciones fueron específicas para las configuraciones proporcionadas, otras fueron genéricas, lo que resalta la necesidad de proporcionar más contexto en ciertos escenarios. El estudio concluye que los modelos de

lenguaje son efectivos para generar recomendaciones accionables en entornos de nube. Sin embargo, se sugiere que futuras iteraciones integren un enfoque híbrido que combine métodos deterministas con análisis basado en lenguaje para mejorar la precisión y la relevancia contextual. También se recomienda ampliar el alcance de la herramienta para incluir otros servicios de GCP.

Palabras clave: Google Cloud Platform, configuraciones de red, VPC, VPN, buenas prácticas, modelos de lenguaje, seguridad en la nube, optimización de redes, análisis híbrido.

Resumen Ejecutivo

Análisis automático de buenas prácticas de arquitectura en implementaciones en Google Cloud Platform

Las configuraciones de redes en Google Cloud Platform (GCP) juegan un papel crucial en la operación de entornos empresariales, ya que afectan directamente la seguridad, eficiencia y mantenibilidad de las infraestructuras en la nube. Sin embargo, la complejidad creciente de estos entornos y la necesidad de cumplir con estándares estrictos presentan desafíos significativos para las organizaciones. En este contexto, el presente trabajo tiene como objetivo desarrollar y evaluar un aplicativo basado en modelos de lenguaje que permita analizar configuraciones de redes en GCP y proponer mejoras fundamentadas en buenas prácticas. Este desarrollo experimental busca facilitar el cumplimiento de estándares de configuración en entornos organizacionales complejos, optimizando aspectos clave como la seguridad, la eficiencia operativa y la sostenibilidad a largo plazo de las redes en la nube.

El proyecto consistió en el desarrollo de un agente automatizado que integra herramientas de Google Cloud y capacidades de procesamiento de lenguaje natural. Este agente recopila configuraciones de recursos en GCP (como redes VPC, VPN, subredes y reglas de firewall) y compara estas configuraciones con las buenas prácticas documentadas por Google. El sistema utiliza un modelo de lenguaje avanzado para analizar las configuraciones y generar un reporte en formato Markdown que incluye. Descripción de las configuraciones actuales: Detalle de redes, túneles VPN, reglas de firewall y subredes encontradas en GCP. Recomendaciones específicas: Análisis basado en la configuración existente. Recomendaciones genéricas: Sugerencias obtenidas de la documentación pública cuando el contexto proporcionado no es suficiente para una evaluación específica.

Para la evaluación del aplicativo, se creó una organización ficticia en GCP (*roma.joonix.net*) con una topología de red representativa que incluyó proyectos de red, proyectos host-spoke y proyectos de servicio, simulando una organización real.

El análisis generó un reporte estructurado con tres secciones principales: a) Configuraciones de red: Se detallaron las redes VPC, subredes, reglas de firewall y túneles VPN. El aplicativo identificó configuraciones específicas que podían simplificarse, como la fusión de redes con nombres similares. Detectó configuraciones inseguras, como una regla de firewall *allow-all* que expone riesgos significativos. b) Resumen de buenas prácticas: Para redes VPC, se incluyeron prácticas relacionadas con nomenclatura, seguridad, conexión entre redes y administración centralizada. En el caso de VPN, se destacaron prácticas sobre alta disponibilidad, enrutamiento, rendimiento y monitoreo. Análisis y recomendaciones. c) Recomendaciones específicas: Algunas sugerencias fueron puntuales, como la necesidad de establecer túneles VPN en regiones separadas para mejorar la conmutación por error. d) Recomendaciones genéricas: Estas incluyeron el uso de descripciones para facilitar la administración, la creación de túneles adicionales para mejorar el rendimiento y la documentación de configuraciones para garantizar la alineación con objetivos organiza-

cionales.

Aunque el aplicativo cumplió con su objetivo principal, se identificaron áreas de mejora, tales como: a) En algunos casos, las recomendaciones eran genéricas porque el modelo no recibió información suficiente, como las configuraciones completas de los extremos de las VPN. b) Combinar el análisis basado en modelos de lenguaje con un análisis determinista podría mejorar la precisión y relevancia de las recomendaciones. El proyecto demostró que los modelos de lenguaje pueden ser herramientas útiles para analizar configuraciones complejas en entornos de nube y proporcionar recomendaciones basadas en buenas prácticas. Sin embargo, es fundamental enriquecer el contexto proporcionado al modelo para maximizar la utilidad de las recomendaciones. Además, un enfoque híbrido podría combinar la flexibilidad de los modelos de lenguaje con la precisión de los análisis deterministas.

Índice de figuras

Índice de tablas

I. Introducción

La computación en la nube ha revolucionado la forma en que las empresas gestionan sus servicios de tecnología de la información, ofreciendo ventajas significativas como la escalabilidad, flexibilidad, disponibilidad, seguridad y ahorro de costos (Saini et al., 2019). Sin embargo, aprovechar al máximo estos beneficios requiere de la implementación de arquitecturas de software adaptadas a las necesidades de cada proyecto. Google Cloud Platform ofrece una variedad de servicios que pueden clasificarse en diferentes categorías según su funcionalidad. En el ámbito del cómputo, proporciona opciones como Google Compute Engine para máquinas virtuales y Google App Engine para aplicaciones web. Para el almacenamiento y gestión de datos, dispone de servicios como Google Cloud Storage para almacenamiento de objetos y Google Cloud Firestore para bases de datos NoSQL. Además, Google Cloud Platform abarca servicios de gestión de contenedores como Google Kubernetes Engine, mensajería con Google Cloud Pub/Sub y un amplio abanico de otras herramientas que se adaptan a las necesidades de desarrollo y gestión de aplicaciones en la nube. Esta diversidad permite a los usuarios elegir las herramientas que mejor se ajusten a sus requerimientos específicos y construir soluciones flexibles y escalables.

A pesar de las ventajas inherentes que ofrece Google Cloud, el diseño y la evaluación de servicios de software en esta plataforma presentan un desafío considerable (Mohammed Sadeeq et al., 2021) (Papadopoulos et al., 2021). Este desafío radica en la necesidad de poseer un profundo conocimiento de los servicios disponibles, sus interacciones y limitaciones para poder seleccionar las herramientas adecuadas, optimizar su configuración, y garantizar que la solución final cumpla con los requisitos de rendimiento, escalabilidad y costo. Sin este conocimiento, es fácil caer en errores de diseño que pueden resultar en un rendimiento deficiente, costos inesperados o incluso la incapacidad de alcanzar los objetivos del proyecto. Aunque Google Cloud cuenta con un servicio de recomendaciones (GCP Recommender, 2023), es esencial destacar que dicho servicio tiene limitaciones notables. Por ejemplo, en el caso del servicio Google Kubernetes Engine (GKE), Google Recommender no aborda aspectos clave, como la conversión de clústeres de Kubernetes al modo de piloto automático, una práctica esencial para optimizar el uso eficiente de recursos y evitar el sobredimensionamiento, y por ende, el sobrecosto de la infraestructura (GCP GKE Configuration Choices, 2023). Otro ejemplo significativo es la omisión de aspectos fundamentales, como el diseño de subredes en Google Cloud, que, a pesar de estar documentado como una buena práctica, no es abordado por el servicio de recomendaciones de Google Cloud (GCP VPC Best practices, 2023). La cobertura de estos vacíos en la funcionalidad de la herramienta GCP Recommender puede tener un valor considerable

tanto en el ámbito académico como profesional.

Otros proveedores líderes de nube pública, como AWS y Azure, también ofrecen servicios similares al Recomendador de Google Cloud, conocidos como AWS Trusted Advisor y Azure Advisor, respectivamente. Sin embargo, es crucial destacar que, al igual que el servicio de Google Cloud, el análisis que realizan estos servicios, se limita a un subconjunto específico de productos (Cass Information Systems, 2023) (Sinha, 2023) (Tozzi, 2023).

La necesidad de avanzar en este campo es evidente, ya que los usuarios de la nube requieren un enfoque más integral para optimizar sus recursos y garantizar la eficiencia y seguridad de sus aplicaciones. En este sentido, este trabajo busca proporcionar soluciones y recomendaciones que complementen el servicio de Google Recommender, enriqueciendo la experiencia de sus usuarios y ampliando las posibilidades de optimización. La fundación de Google Cloud es un componente clave de arquitectura y la comunicación interna y externa requiere un análisis que permita a los usuarios adoptar las mejores prácticas de integración mediante redes de VPC y redes de VPN.

II. Descripción General

Oportunidad y problemática

En el contexto del servicio de Google Recommender, la oportunidad surge de la observación de que dicho servicio no abarca todas las buenas prácticas descritas en la documentación de Google Cloud o, en su defecto, los usuarios no revisan exhaustivamente dicha documentación (Papadopoulos et al., 2021). Esta situación conlleva a una problemática importante: la subutilización de las ventajas y funcionalidades ofrecidas por la nube de Google Cloud, lo cual puede limitar la eficiencia operativa y el rendimiento de las aplicaciones desplegadas en esta plataforma. En este contexto, se presenta una valiosa oportunidad para cerrar esta brecha mediante el desarrollo de un software capaz de analizar automáticamente las configuraciones de infraestructura en Google Cloud y proporcionar recomendaciones específicas al usuario sobre cómo mejorar y adoptar las buenas prácticas recomendadas por la plataforma (Tozzi, 2023). Este enfoque no solo busca optimizar el uso de los recursos en la nube, sino también fomentar una mayor adopción de las prácticas recomendadas para garantizar un entorno de computación en la nube eficiente y seguro.

El objetivo principal de este proyecto es la identificación de servicios en Google Cloud que, a pesar de contar con documentación de buenas prácticas, requieren comprobaciones manuales adicionales no cubiertas por Google Cloud Recommender (GCP Recommender, 2023). Una vez identificados estos servicios, se llevará a cabo una exhaustiva evaluación de lenguajes, técnicas y tecnologías, con el fin de desarrollar un sistema de análisis automático de la infraestructura en Google Cloud.

Se identificarán aquellas mejores prácticas en ciertos servicios seleccionados que carezcan de comprobaciones automáticas en Google Cloud. Para cada una de estas prácticas, se crearán reglas detalladas que permitirán evaluar su cumplimiento. Estas reglas se desarrollarán de manera rigurosa durante la ejecución del proyecto y **se basarán en la documentación oficial de Google Cloud**, garantizando así la objetividad y la validez de las recomendaciones.

Desde una perspectiva de **requisitos funcionales de alto nivel**, la herramienta requerirá acceso a la consola y a la organización de Google Cloud que se escaneará (El término **organización** hace referencia a una estructura de recursos creada en Google Cloud). En primer lugar, la herramienta ejecutará un análisis a través de Google Recommender (GCP Recommender, 2023). Luego, realizará un análisis complementario basándose en las soluciones y reglas definidas previamente, con el fin de generar un informe integral de

recomendaciones basado en los productos utilizados por el usuario en Google Cloud. Estos informes se almacenarán en el servicio de almacenamiento de datos de Google Cloud, conocido como BigQuery (GCP BigQuery, 2023). Posteriormente, los resultados se visualizarán a través del servicio de datos de Google llamado Looker Studio (GCP Looker Studio, 2023), lo que permitirá a los usuarios observar de manera efectiva los resultados de cada análisis y tomar medidas para corregir las configuraciones que no cumplan con las buenas prácticas.

Es importante destacar que el desarrollo de esta herramienta **requiere una cuenta de facturación de Google Cloud** para cubrir los costos asociados con el uso de servicios como Google Recommender (GCP Recommender, 2023), BigQuery (GCP BigQuery, 2023) y Looker Studio (GCP Looker Studio, 2023). No obstante, para este proyecto en particular, no se incurrirán costos directos, ya que se utilizará una organización de prueba proporcionada por Google. Como gesto de agradecimiento a Google por proporcionar este entorno de prueba sin costo, se planea iniciar el proceso de publicación de la herramienta en un repositorio público de herramientas complementarias para Google Cloud (GCP Github Repository, 2023), lo que enriquecerá el conjunto de recursos disponibles para los clientes de Google Cloud. Es importante señalar que los costos asociados con el uso continuo de la herramienta una vez que esté publicada **serán responsabilidad de los usuarios**.

En última instancia, el objetivo de este proyecto es reducir los errores humanos, la subjetividad y las interpretaciones incorrectas de configuraciones, problemas comunes que a menudo resultan de la falta de experiencia o de sesgos basados en prácticas antiguas (Menzies and Zimmermann, 2013) (Menzies and Zimmermann, 2018). La herramienta proporcionará una solución objetiva y automatizada para evaluar y mejorar las configuraciones en Google Cloud, beneficiando a los usuarios y optimizando de manera efectiva sus recursos.

III. Descripción del proyecto

Objetivo General

Desarrollar una herramienta de análisis automatizado de infraestructura en Google Cloud con el fin de respaldar la adopción de las mejores prácticas proporcionadas por esta tecnología.

Objetivos Específicos

1. Seleccionar los productos y características de Google Cloud que aún no cuentan con la revisión automática de configuraciones de buenas prácticas.
2. Diseñar una herramienta para el análisis automático de infraestructura para los servicios seleccionados anteriormente.
3. Implementar la herramienta diseñada mediante el uso de tecnologías compatibles con el API de Google Cloud.
4. Evaluar el funcionamiento de la herramienta propuesta mediante un estudio de caso de una arquitectura de Google Cloud

Fases de desarrollo

El trabajo de grado se divide en cuatro fases alineadas con los objetivos específicos mencionados anteriormente (ver figura **III-1**). Con más de cien servicios disponibles en Google Cloud (GCP Service Summary, 2023), la primera fase se centra en seleccionar servicios con documentación de buenas prácticas pero sin análisis automático de infraestructura. Se definirán criterios de selección para esta elección. Cabe notar que el **alcance** de las siguientes fases será exclusivo para los servicios seleccionados en esta fase, debido a limitaciones de tiempo y recursos.

La segunda fase de este proyecto se centra en el diseño de una herramienta de análisis automático para los servicios previamente seleccionados. Inicialmente, se llevarán a cabo pruebas de concepto para evaluar la compatibilidad de tecnologías, bibliotecas y lenguajes con Google Cloud. Basándonos en los resultados

Figura **III-1**.: Fases metodológicas del trabajo de grado.

de estas pruebas, se diseñará la herramienta de análisis automático de buenas prácticas. En esta fase, se pretende establecer un diseño basado en la documentación de buenas prácticas. Por ejemplo, si uno de los productos seleccionados en la fase anterior fue BigQuery (GCP BigQuery Best Practices, 2023), se examinará qué aspectos de la documentación de buenas prácticas son cubiertos por el Recomendador de Google Cloud (GCP Recommender, 2023). Para aquellos aspectos que no están cubiertos, se identificarán los que pueden ser verificados de manera automática, y se generará una estrategia para cada uno de ellos.
La tercera fase se enfocará en la implementación de la herramienta de análisis automático, siguiendo un enfoque de desarrollo en cascada. Esto incluirá la planificación, diseño, desarrollo y pruebas de la implementación. Además, se diseñarán los mecanismos necesarios para procesar documentación futura, lo que permitirá extender la funcionalidad de la herramienta. La herramienta resultante, desarrollada en esta fase, será utilizada en la cuarta fase y se someterá a un proceso de publicación en el repositorio de herramientas complementarias de Google Cloud. Es importante tener en cuenta que para el funcionamiento de esta herramienta, se requerirá acceso a una organización de Google Cloud con una cuenta de facturación habilitada que asuma los costos asociados.
La fase 4 busca validar la herramienta mediante un estudio de caso. Se seleccionará y describirá un estudio de caso, proporcionando un contexto detallado para las pruebas. Se presentarán resultados, conclusiones y propuestas para futuros trabajos.
El estudio propuesto adopta un enfoque metodológico cuantitativo, fundamentado en la investigación aplicada según la perspectiva de (Bryman, 2016). Este enfoque implica la utilización de métodos y técnicas que permiten recopilar y analizar datos de manera sistemática y objetiva, centrándose en la aplicación práctica de los conocimientos teóricos en un contexto específico. En este caso, la investigación se enfoca en el ámbito de la computación en la nube, particularmente en el entorno de Google Cloud.
La metodología se apoya en el uso de herramientas de autenticación y autorización proporcionadas por Google Cloud, las cuales permiten acceder de manera segura a los recursos y servicios de la plataforma. Estas herramientas desempeñan un papel crucial en la validación de la identidad de los usuarios y en la gestión de los permisos de acceso, garantizando así la confidencialidad, integridad y disponibilidad de los datos y sistemas en la nube.
Además, se emplearán herramientas especializadas de análisis de configuraciones en Google Cloud las cuales se usarán en la tercera fase. Estas herramientas están diseñadas para evaluar de manera detallada las configuraciones de infraestructura, identificar posibles vulnerabilidades o desviaciones de las buenas prácticas establecidas por Google, y proporcionar recomendaciones específicas para mejorar la seguridad, eficiencia y rendimiento de las arquitecturas en la nube.
Al seguir este enfoque metodológico riguroso y sistemático, se espera diseñar una herramienta efectiva y práctica que contribuya significativamente a la optimización y seguridad de las arquitecturas en Google Cloud, proporcionando un valor agregado tanto para los usuarios como para las organizaciones que operan en entornos de nube pública.

IV. Marco teórico y trabajos relacionados

IV.1. Computación en la nube

La definición de la computación en la nube varía, pero el Instituto Nacional de Estándares y Tecnología (NIST) la conceptualiza como “un modelo que permite el acceso a la red de manera ubicua, conveniente y bajo demanda a un conjunto compartido de recursos computacionales configurables”(Mell and Grance, 2011). Estos recursos, que abarcan desde redes y servidores hasta almacenamiento, aplicaciones y servicios, pueden ser aprovisionados y liberados de manera rápida y eficiente con un esfuerzo mínimo de gestión o interacción con el proveedor de servicios (Furht and Escalante, 2010) .

La computación en la nube representa un paradigma en el cual se integran servicios mediante la utilización de recursos a través de Internet, y se escalan de manera dinámica. Inicialmente, el término “nube”solía denotar una porción de la infraestructura de Internet. Sin embargo, en la actualidad, la noción de la nube ha evolucionado, actuando como una metáfora para describir servicios entregados a través de Internet. La rápida progresión de los servicios en la nube ha posibilitado la ejecución de numerosas operaciones en fracciones de segundo, contrastando con los sistemas tradicionales que tenían limitaciones en la cantidad de transacciones que podían manejar (Surbiryala and Rong, 2019). Este poder computacional recién adquirido es aplicable a una amplia variedad de tareas, que abarcan desde el preprocesamiento hasta el análisis y la predicción de eventos futuros. Los servicios en la nube han transformado la naturaleza de las operaciones, permitiendo a los usuarios realizar tareas complejas de manera eficiente y en tiempo real (Surbiryala and Rong, 2019). Para aquellos lectores interesados en profundizar en los conceptos relacionados con la nube, se recomienda consultar el anexo A, donde se ofrece una exploración más detallada de estos temas.

IV.2. Modelos de servicio de nube pública

Los servicios de nube pública agilizan el proceso de transformación tecnológica, no obstante cabe notar que los modelos de servicio difieren unos de otros dependiendo de las responsabilidades del usuario final. Los tres grandes tipos son el de Infraestructura como un Servicio (IaaS), Plataforma como un servicio (PaaS) y Software como un servicio (SaaS), no obstante en esta sección se hace referencia a tipos más granulares de servicio como Contenedores como un servicio, Funciones como un servicio .

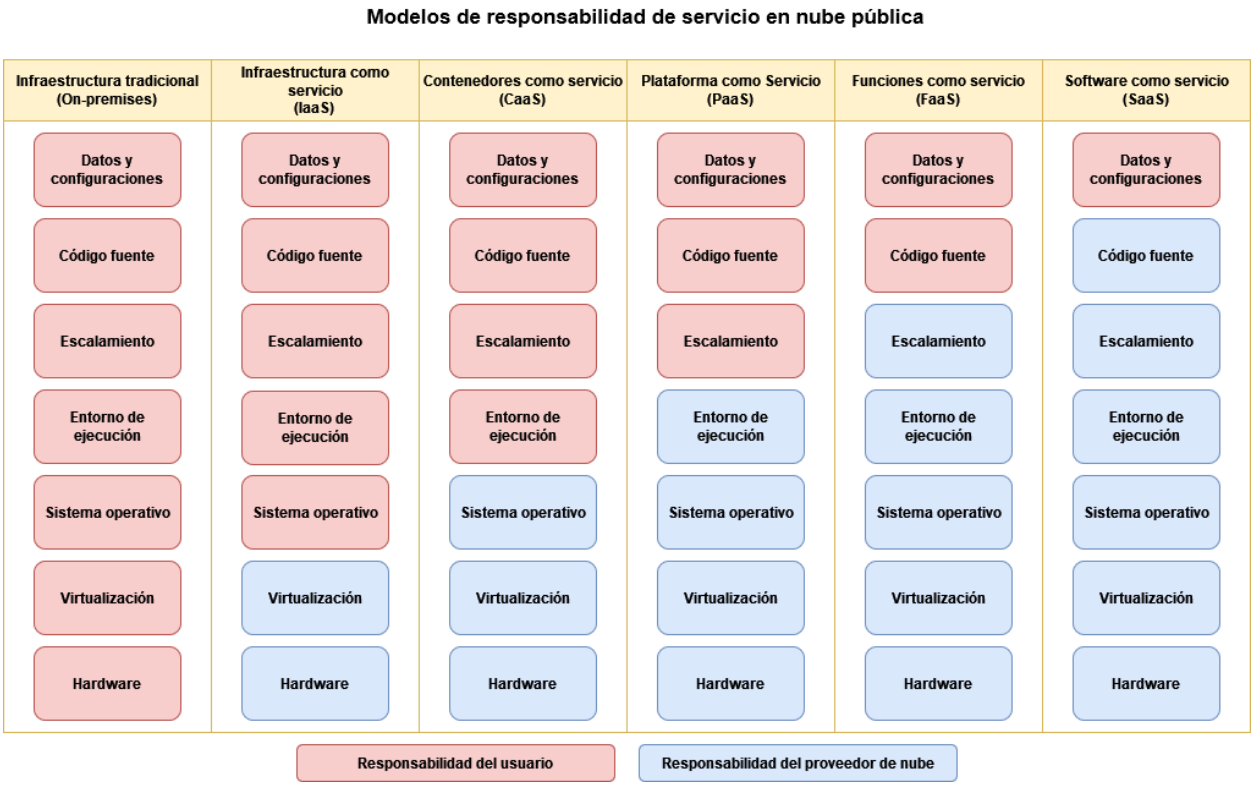


Figura **IV-1**.: Modelos de responsabilidad de servicio en nube pública. Adaptado de (Google Service Models, 2024)

En la Figura **IV-1**, en primer lugar, se presenta el modelo de infraestructura tradicional, donde las empresas y usuarios son responsables de la adquisición, administración y aprovisionamiento del hardware, así como de todos los componentes de configuración necesarios para ofrecer servicios de software a los usuarios finales. Este enfoque implica que las empresas deben proporcionar no solo el hardware, sino también el espacio físico, las condiciones energéticas, los sistemas de enfriamiento, el mantenimiento físico y de redes, lo que agrega una carga operativa significativa. Esto puede resultar desventajoso, especialmente cuando el enfoque principal de las empresas no está en el mantenimiento de la infraestructura tecnológica.

Posteriormente, en la Figura **IV-1**, se ilustra el modelo de Infraestructura como Servicio (IaaS), donde la responsabilidad del hardware y la virtualización pasa del cliente al proveedor de la nube. Sin embargo, la actualización, configuración y mantenimiento del sistema operativo siguen siendo responsabilidad de los clientes, lo que aún implica un esfuerzo considerable por parte de las empresas. Lo mismo ocurre con el servicio de Contenedores como Servicio, donde, a diferencia del servicio de infraestructura, solo el sistema operativo deja de ser responsabilidad de los clientes y pasa a ser responsabilidad del proveedor de la nube.

A partir de los servicios de Plataforma como Servicio (PaaS), Funciones como Servicio (FaaS) y Software como Servicio (SaaS), los usuarios de la nube pueden prescindir de los detalles de la infraestructura y con-

centrarse en elementos operativos como el código fuente, los datos y las configuraciones. Estos aspectos están más relacionados con la administración de los servicios de software que con la administración de servicios de infraestructura, como se muestra en la Figura **IV-1**.

Los proveedores de nube pública ofrecen una variedad de productos adaptados a cada modelo de servicio, lo que permite a los usuarios elegir entre opciones más nativas de la nube o enfoques más tradicionales. Por consiguiente, resulta crucial comprender que cada uno de estos productos conlleva directrices y buenas prácticas específicas que deben ser adoptadas para maximizar su uso.

Esta comprensión profunda y la adopción de las mejores prácticas son fundamentales para garantizar la eficacia y la eficiencia en el aprovechamiento de los servicios en la nube pública. Al seguir los lineamientos recomendados por el proveedor y adoptar las prácticas óptimas para cada producto, los usuarios pueden optimizar su infraestructura y maximizar los beneficios que ofrece la nube pública.

IV.2.1. Proceso de Fundación de Infraestructura de Google Cloud

El proceso de fundación de infraestructura en Google Cloud (Google Cloud Setup, 2024) se refiere a la planificación, diseño, implementación y configuración inicial de los recursos y servicios necesarios para establecer un entorno de nube funcional. Este proceso incluye la definición de redes, políticas de seguridad, estructura de cuentas, gestión de identidades, y acceso, así como la configuración de herramientas de monitoreo y cumplimiento. La fundación de infraestructura es crítica porque establece las bases para todas las operaciones futuras en la nube. **Escalabilidad**: Proveer una arquitectura capaz de crecer con las necesidades de la organización. **Seguridad**: Implementar controles que protejan datos y aplicaciones. **Eficiencia operativa**: Garantizar que los recursos se utilicen de manera óptima. **Cumplimiento**: Facilitar la adherencia a normativas y estándares corporativos. Conectividad: Asegurar una integración efectiva entre los sistemas en la nube y los entornos on-premises.

IV.2.2. Redes de VPC y VPN en el marco de Google Cloud

Una Red de VPC (Virtual Private Cloud) (Google Cloud VPC, 2024) en Google Cloud funciona como una red virtual global que facilita la conexión y comunicación privada de los recursos dentro de una organización. Al abarcar subredes en múltiples regiones, cada VPC otorga un control total sobre aspectos esenciales como el direccionamiento IP, las rutas, las reglas de firewall y la conectividad híbrida. Esto permite segmentar el tráfico, restringir accesos mediante reglas de firewall y políticas, y proveer entornos separados para diferentes aplicaciones o equipos dentro de la organización. Además, las redes VPC ofrecen soporte para configuraciones personalizadas, escalabilidad global y facilitan la interconexión entre recursos en Google Cloud, otras nubes y entornos on-premises mediante peering o VPN. De esta manera, contribuyen al cumplimiento de requisitos regulatorios a través de la separación de redes y la monitorización de tráfico.

Por otro lado, la conectividad híbrida de VPN (Virtual Private Network) en Google Cloud establece una conexión segura entre las redes locales (on-premises) y las redes virtuales en la nube. Mediante protocolos de cifrado, protege los datos en tránsito y garantiza la comunicación privada entre ambas ubicaciones. Esta conectividad es esencial para facilitar la coexistencia de recursos en la nube y en las instalaciones locales, permitiendo una migración gradual o una operación híbrida permanente. Además de proporcionar una conexión cifrada para proteger datos sensibles, permite a las organizaciones expandir sus capacidades sin depender únicamente de soluciones en la nube o en local. Asimismo, la conectividad híbrida de VPN garantiza el acceso a recursos críticos en diferentes ubicaciones en caso de fallos o interrupciones, optimizando la transferencia de datos al reducir la latencia y mejorar la estabilidad mediante rutas dedicadas o configuraciones de alta disponibilidad.

IV.3. LLMs - Modelos de lenguaje extensos

Los grandes modelos de lenguaje, comúnmente conocidos como **LLMs** (por sus siglas en inglés, *Large Language Models*), son sistemas de inteligencia artificial diseñados para procesar y generar texto humano a partir de enormes cantidades de datos y arquitecturas avanzadas de redes neuronales. Su desarrollo ha sido impulsado por avances en el aprendizaje profundo, en particular en las arquitecturas de redes neuronales transformadoras (*transformers*) que permiten analizar y relacionar grandes volúmenes de información textual para producir respuestas coherentes, contextuales y altamente detalladas.

El entrenamiento de un LLM implica procesar miles de millones de palabras para aprender patrones lingüísticos, relaciones semánticas, sintaxis y el contexto en el lenguaje. Durante este proceso, el modelo analiza las asociaciones y dependencias entre palabras en grandes conjuntos de datos, que abarcan múltiples dominios, incluyendo ciencia, literatura, cultura y datos técnicos. La arquitectura de los modelos LLM se basa generalmente en capas de transformadores, una tecnología que utiliza la atención auto-regresiva (Lütkepohl, 2013) (Teräsvirta, 1994) para considerar la relación entre cada palabra en una oración y todas las demás palabras. Esto permite al modelo captar el contexto y las dependencias de una oración, de manera que pueda generar texto con coherencia gramatical y contextual.

El proceso de entrenamiento requiere de grandes cantidades de poder computacional y recursos de procesamiento gráfico, ya que el modelo pasa por millones de parámetros que optimizan su rendimiento (Patel et al., 2024). Estos parámetros son pesos y sesgos que se ajustan en función de la relación entre palabras y el contexto en que aparecen. Durante el entrenamiento, los modelos también se exponen a diferentes estilos de escritura y fuentes de datos, lo cual permite que desarrollen una comprensión del lenguaje que es tanto genérica como especializada, dependiendo de las necesidades del usuario.

La característica distintiva de los LLMs es su capacidad para comprender y generar lenguaje de manera contextualmente relevante, adaptándose a la intención del usuario. Al recibir una entrada de texto, el mo-

delo evalúa no solo las palabras, sino también el contexto y las implicaciones subyacentes, permitiéndole ofrecer respuestas que son más precisas y coherentes. Esta comprensión contextual es posible gracias a los mecanismos de atención en los transformadores, los cuales permiten que el modelo focalice recursos computacionales en las partes más relevantes de un texto de entrada (Briganti, 2024).

El uso de LLMs se ha expandido rápidamente en diversos campos, desde el procesamiento de lenguaje natural en chatbots hasta la creación de contenido, traducción automática, análisis de sentimientos y generación de código. En entornos empresariales, los LLMs están facilitando la automatización de tareas repetitivas y de procesamiento de texto, permitiendo que los empleados se enfoquen en labores de mayor valor agregado. Por ejemplo, en el sector legal, los modelos de lenguaje pueden analizar grandes volúmenes de documentos (Fagan, 2024), mientras que en marketing pueden generar descripciones de productos o análisis de audiencias ().

IV.4. Agentes de LangChain

Los agentes de LangChain (LangChain, 2024) son componentes sofisticados diseñados para manejar de forma dinámica tareas de procesamiento de lenguaje natural (NLP) mediante grandes modelos de lenguaje (LLMs) y otras herramientas. A diferencia de los modelos de lenguaje tradicionales que requieren entradas específicas y limitadas, los agentes de LangChain ofrecen un enfoque más flexible y adaptable. Pueden recibir preguntas o instrucciones amplias y determinar cómo desglosarlas en subtareas, identificar qué herramientas adicionales necesitan para resolver el problema, y generar respuestas precisas y contextualizadas. Estos agentes permiten el uso de LLMs no solo como generadores de texto sino como componentes autónomos capaces de seguir flujos de trabajo complejos y de interactuar con otros recursos de software y datos.

En el corazón de un agente de LangChain está un proceso de razonamiento iterativo que lo convierte en una herramienta poderosa para resolver problemas de varias etapas o que requieren una consulta de recursos externos. Al recibir una solicitud, el agente examina el contexto y el tipo de tarea, lo que le permite decidir de forma autónoma si la información disponible en el modelo es suficiente o si se requiere apoyo de herramientas externas (LangChain, 2024). Los agentes de LangChain son particularmente útiles cuando se trabaja con múltiples fuentes de datos o herramientas de soporte, como bases de datos, APIs específicas, o incluso otros modelos de lenguaje.

Un agente de LangChain puede, por ejemplo, desglosar una consulta compleja en preguntas más específicas o acceder a herramientas como navegadores para obtener información actualizada que no está en el modelo. A continuación, integra las respuestas obtenidas en un solo flujo lógico y continuo. Si el agente necesita consultar una API externa, extrae la información relevante, la procesa y luego formula una respuesta compuesta que considera todos los datos y análisis involucrados. Este enfoque iterativo permite a

los agentes mantener un alto nivel de precisión y contextualización, sin importar la complejidad de la tarea (Topsakal and Akinci, 2023).

Los agentes de LangChain también son particularmente efectivos para tareas que requieren herramientas adicionales fuera del propio modelo de lenguaje. Al estar diseñados con una arquitectura flexible, los agentes pueden conectarse a herramientas externas como bases de datos SQL, motores de búsqueda, APIs web, e incluso entornos de ejecución de código en Python para obtener resultados en tiempo real o realizar cálculos complejos (Nascimento et al., 2023). Por ejemplo, si un usuario hace una consulta que involucra operaciones matemáticas avanzadas o preguntas sobre información actualizada que no está almacenada en el modelo, el agente puede invocar un intérprete de Python o una API de búsqueda, extraer la información y luego integrar estos datos con sus propios análisis (Soygazi and Oguz, 2023).

Cuando se conectan a estas herramientas, los agentes no solo “llaman” a estas funciones; en realidad, configuran una serie de instrucciones y verificaciones para asegurarse de que las respuestas obtenidas sean precisas y relevantes para la tarea en curso. A través de este proceso de evaluación y verificación, los agentes pueden minimizar el margen de error y garantizar que las respuestas finales sean completas y contextualmente adecuadas (LangChain, 2024).

Uno de los aspectos más innovadores de los agentes de LangChain es su capacidad de razonamiento iterativo. En lugar de proporcionar una respuesta inmediata a cada pregunta, los agentes pueden llevar a cabo un proceso iterativo de prueba y ajuste en el cual evalúan si su respuesta es adecuada o si necesitan realizar más consultas. Esto es especialmente útil en consultas amplias o en aquellas que requieren análisis multi-etapa (Soygazi and Oguz, 2023). Por ejemplo, si se formula una pregunta sobre la comparación de estrategias de inversión, un agente puede comenzar con una consulta inicial, obtener una respuesta preliminar, analizar esa respuesta y luego realizar consultas adicionales para aclarar o expandir la información hasta que esté seguro de haber proporcionado una respuesta exhaustiva.

Este proceso iterativo permite a los agentes de LangChain adaptarse en tiempo real a la información que van obteniendo, mejorando su comprensión a medida que analizan los datos. Los agentes revisan cada respuesta parcial y la comparan con la consulta original, determinando si el resultado actual responde de manera completa a la pregunta del usuario o si es necesario profundizar en algún aspecto específico. Este enfoque proporciona un mayor grado de precisión en las respuestas finales, al mismo tiempo que demuestra una capacidad para manejar problemas complejos o ambivalentes (Easin Arafat et al., 2023).

Este tipo de aplicaciones demuestra la capacidad de los agentes para no solo generar texto basado en una entrada, sino realizar una serie de operaciones complejas, accediendo a fuentes de información variadas y usando lógica iterativa para llegar a conclusiones más precisas.

IV.5. LangGraph

LangGraph (LangGraph, 2024) es una extensión reciente dentro del ecosistema de LangChain que busca mejorar la orquestación de flujos de trabajo complejos en aplicaciones de IA generativa. Mientras que LangChain permite construir aplicaciones con modelos de lenguaje de gran escala (LLMs) integrando de manera modular múltiples herramientas y agentes, LangGraph añade la capacidad de organizar estos procesos en grafos de tareas más avanzados y flexibles.

LangGraph es un marco de trabajo que introduce la noción de ”flujos de trabajo como grafos” en la integración con LLMs (Jeong, 2024). Un grafo en este contexto es una estructura compuesta por nodos y aristas, donde cada nodo representa una tarea o acción específica (por ejemplo, una consulta de API, un análisis de datos o una interacción con el usuario) y las aristas representan la dependencia entre tareas. LangGraph permite definir estos flujos de trabajo de una manera visual y escalable, de modo que cada paso del proceso puede depender de múltiples entradas y generar múltiples salidas.

LangGraph se utiliza para diseñar flujos de trabajo complejos que requieren tomar decisiones o activar tareas en función de resultados anteriores (Jeong, 2024). Con esta herramienta, un desarrollador puede definir, por ejemplo, un proceso que: Obtenga información de distintas fuentes, Decida cuál herramienta o modelo de IA debe emplearse en función de esos datos, Ejecute las tareas en paralelo o en secuencia, de acuerdo con las dependencias definidas.

Esto es especialmente útil para aplicaciones donde el flujo de trabajo tiene pasos condicionales o bifurcaciones, como en el análisis de datos en entornos empresariales, la automatización de consultas o la generación de contenido que dependa de varias fuentes de datos o contextos (Lin et al., 2024).

LangChain proporciona el núcleo de las capacidades de interacción con los modelos de lenguaje, mientras que LangGraph organiza estos componentes en una estructura de grafo que facilita su ejecución. En esencia: **LangChain** gestiona la interacción con LLMs, agentes y herramientas, permitiendo integrar modelos y ejecutar tareas programáticamente. **LangGraph** organiza y administra la secuencia y dependencia de esas tareas, convirtiéndolas en nodos dentro de un flujo de trabajo estructurado.

Al combinar ambas, LangGraph facilita la creación de aplicaciones más sofisticadas en LangChain, permitiendo definir procesos de varias etapas con lógica condicional y bifurcaciones, todo dentro de un marco estructurado y escalable.

IV.6. LangSmith

LangSmith (LangSmith , 2024) es una herramienta diseñada para el desarrollo, monitoreo y depuración de aplicaciones impulsadas por agentes de inteligencia artificial. Su propósito es ayudar a los desarrolladores a gestionar y mejorar los modelos de lenguaje y flujos de trabajo de IA, especialmente aquellos que interactúan de manera iterativa o realizan tareas complejas.

Con LangSmith, los desarrolladores pueden rastrear el rendimiento de los agentes, analizar sus decisiones y optimizar su flujo de trabajo. Es especialmente útil en proyectos construidos con LangChain, ya que facilita la supervisión y mejora continua de los modelos, contribuyendo a su precisión y efectividad en aplicaciones de procesamiento de lenguaje natural (Ito et al., 2020).

IV.7. Trabajos relacionados

En la sección previa (Sección IV.2), se detallan los diversos modelos de servicio ofrecidos por la nube pública, los cuales son facilitados por una amplia gama de productos disponibles en múltiples proveedores de servicios en la nube. En comparación con el modelo de infraestructura tradicional en las instalaciones (on-premises), la computación en la nube representa una ventaja significativa en términos de costos operativos y de mantenimiento de la tecnología necesaria para el funcionamiento de cualquier empresa.
Sin embargo, la diversidad de modelos de servicio, proveedores de nube y productos tecnológicos plantea un desafío al momento de utilizar la computación en la nube para gestionar la infraestructura de manera eficiente. Este desafío ha sido reconocido por la comunidad académica, la cual ha realizado contribuciones significativas a través de diversos trabajos relacionados con la optimización y gestión de la infraestructura en la nube.

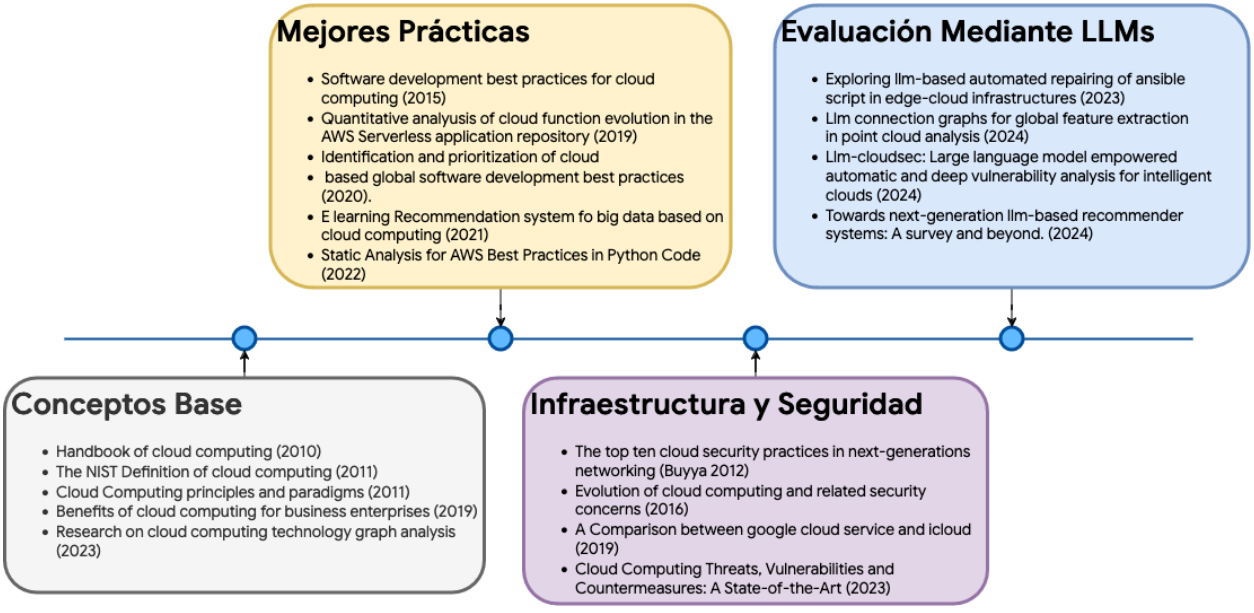


Figura **IV-2**.: Trabajos relacionados y contribuciones académicas a la fundación de conceptos, establecimiento de buenas prácticas de desarrollo, infraestructura y seguridad para la computación en la nube, evaluación mediante LLMs.

Como se ilustra en la Figura **IV-2**, tras el lanzamiento inicial de los servicios de Software como Servicio

(SaaS) e Infraestructura como Servicio (IaaS), surgieron diversas contribuciones significativas en el ámbito de la computación en la nube. Entre ellas se destacan el “Handbook of Cloud Computing” (Furht and Escalante, 2010) , definiciones clave proporcionadas por gigantes tecnológicos como Oracle (Mell and Grance, 2011), así como principios y paradigmas fundamentales de la computación en la nube (Buyya et al., 2011). Además, se realizaron análisis exhaustivos de los potenciales beneficios empresariales derivados de la adopción de la computación en la nube, junto con investigaciones detalladas sobre los servicios disponibles en diversas nubes públicas, representados mediante gráficos de redes (Ren et al., 2023).

Desde aproximadamente el año 2015, se ha observado un esfuerzo sostenido para establecer lineamientos y mejores prácticas en el desarrollo de software para la nube, con el objetivo de aprovechar al máximo las ventajas ofrecidas por los servicios de nube pública. Asimismo, se han emprendido iniciativas para realizar análisis cuantitativos y priorizar las mejores prácticas, basándose en el establecimiento de estándares de calidad. Además, se han desarrollado métodos para automatizar la detección de buenas prácticas, como se evidencia en el análisis estático de implementaciones en Python desplegadas en AWS (Mukherjee et al., 2022), mencionado en la sección de mejores prácticas de la Figura **IV-2**.

La popularización de los servicios en la nube ha conllevado una creciente preocupación por la seguridad y las configuraciones de infraestructura. Como resultado, se han realizado numerosas contribuciones destinadas a establecer buenas prácticas en estos ámbitos, así como a comparar diversos servicios de Software como Servicio (SaaS), como en el caso de Google Service versus iCloud (Arif et al., 2019). Por ejemplo, el trabajo mencionado recopila el estado del arte de diversas publicaciones relacionadas con amenazas y vulnerabilidades, abordando específicamente 183 trabajos relacionados exclusivamente en el área de seguridad (Pericherla, 2023).

En el contexto de la computación en la nube, la popularización y el lanzamiento de los grandes modelos de lenguaje (LLMs, por sus siglas en inglés) han impulsado avances significativos en su aplicación. Por ejemplo, el uso de grafos basados en LLMs permite la extracción de características específicas de la infraestructura en la nube, facilitando análisis más detallados y efectivos (Wang et al., 2024b). Estas características pueden emplearse para identificar vulnerabilidades en configuraciones de la nube (Cao and Jun, 2024) o para analizar y reparar automáticamente archivos de configuración, como los utilizados en herramientas de automatización tipo Ansible (Kwon et al., 2023).

Además, el análisis de configuraciones es un campo que ha comenzado a beneficiarse notablemente del poder de los LLMs. Tal como lo destaca (Wang et al., 2024a), estos modelos tienen la capacidad de potenciar sistemas de recomendación y análisis, llevando estas herramientas a una nueva generación mediante su integración. Esto abre un enorme potencial para que los LLMs transformen la gestión, optimización y seguridad de los entornos en la nube.

Cabe señalar que estos desarrollos son relativamente recientes, con publicaciones mayoritariamente a partir del año 2023. Esto evidencia el auge de investigaciones y aplicaciones en torno a los LLMs, subrayando la relevancia de contribuir a la exploración de sus capacidades para mejorar evaluaciones y configuraciones

en la nube.

Es importante tener en cuenta que esta sección de trabajos relacionados proporciona un panorama general de las contribuciones significativas en el campo de la computación en la nube, destacando la relevancia y pertinencia del presente trabajo. Sin embargo, no pretende ser una lista exhaustiva de todas las contribuciones realizadas hasta la fecha en este ámbito en constante evolución.

1. Análisis de buenas prácticas en Google Cloud

1.1. Análisis actual de buenas prácticas en Google Cloud

En esta sección, se aborda el análisis del estado actual y la selección de servicios, tal como se describe en la fase 1 de la metodología (ver sección III). Como se ha mencionado a lo largo del documento, Google Cloud ofrece un servicio denominado Google Cloud Recommender (GCP Recommender, 2023), que proporciona recomendaciones e insights sobre la configuración de algunos de sus productos. El objetivo de esta sección es, por lo tanto, enumerar los productos que cuentan con recomendaciones en Google Cloud Recommender y revisar la documentación de mejores prácticas para comprender el alcance actual del servicio.
La Tabla **1-1** contiene los servicios que, según la documentación (GCP Recommender, 2023), disponen de recomendaciones en Google Cloud. Un recommender es un servicio de Google Cloud que proporciona sugerencias de uso para los recursos dentro de la plataforma. Cada recomendador se enfoca en un producto y tipo de recurso específico dentro de Google Cloud. Un mismo producto puede tener varios recomendadores, cada uno ofreciendo recomendaciones diferentes para distintos recursos. Los campos de la tabla incluyen: **Servicio:** el servicio de Google Cloud al cual se aplican las recomendaciones. **Categoría:** Costo, Seguridad, Rendimiento, Fiabilidad, Mantenibilidad, según el tipo de optimización que ofrece la recomendación. textbfNombre: el nombre abreviado del recomendador. **Descripción:** una breve explicación de la función del recomendador.

Tabla **1-1**.: Conjunto de recomendadores existentes para Google Cloud a la fecha 20 de noviembre de 2024

Servicio	Categoría	Nombre	Descripción
BigQuery	Costo	Recomendación de uso de *"Slot"*	Optimiza el costo del uso de los *"slot"* por medio de recomendaciones.
BigQuery	Costo	Recomendador de particiones y/o clústeres	Ayuda a reducir el costo del servicio mediante clusterización o particionamiento de los datos en BigQuery.
Cloud Run	Costo	Asignación de recursos de CPU	El recomendador analiza automáticamente el tráfico recibido por tu servicio de Cloud Run durante el último mes y recomendará cambiar de asignación de CPU durante las solicitudes a asignación permanente de CPU, si esto resulta más económico.
Cloud Run	Seguridad	Recomendador de ajustes de seguridad para Cloud Run	El Recomendador aumenta la seguridad al optimizar: Las cuentas de servicio para un servicio de Cloud Run para que tengan el conjunto mínimo de permisos requeridos. La seguridad de los siguientes elementos en variables de entorno: Contraseñas, Claves API,Credenciales de Aplicación de Google

Servicio	Categoría	Nombre	Descripción
Cloud SQL	Costo	Recomendador de instancias inactivas	El recomendador de instancias inactivas de Cloud SQL analiza las métricas de uso de las instancias principales que tienen más de 30 días de antigüedad.
Cloud SQL	Costo	Recomendador de sobreaprovisionamiento de recursos	El recomendador de sobreaprovisionamiento de Cloud SQL analiza las métricas de uso de las instancias principales que tienen más de 30 días de antigüedad. Para cada instancia, el recomendador considera la utilización de CPU y memoria basándose en los valores de ciertas métricas dentro de los últimos 30 días. El recomendador no analiza réplicas de lectura.
Cloud SQL	Rendimiento	Recomendador de rendimiento de Cloud SQL.	Mejora el rendimiento de las instancias de Cloud SQL: MySQL: aumenta el tamaño de la caché de tablas. MySQL: gestiona un alto número de tablas. PostgreSQL: evita el envolvimiento del ID de transacción.
Cloud SQL	Rendimiento	Recomendador de instancia subdimensionada de Cloud SQL.	El recomendador de instancias subdimensionadas te ayuda a detectar instancias con alta utilización de CPU y/o memoria. Luego, proporciona recomendaciones sobre cómo optimizar la instancia. Esta página describe cómo funciona este recomendador y cómo usarlo.
Cloud SQL	Fiabilidad	Recomendador de confiabilidad de Cloud SQL.	El recomendador de habilitación de alta disponibilidad de Cloud SQL genera recomendaciones de forma proactiva que te ayudan a cumplir con los Acuerdos de Nivel de Servicio (SLA) al proporcionar redundancia de datos en tus instancias importantes. Esto puede ser útil durante un corte zonal o cuando una instancia se queda sin memoria..
Cloud SQL	Fiabilidad	Recomendador de falta de espacio en disco de Cloud SQL	El recomendador de falta de espacio en disco de Cloud SQL genera recomendaciones de forma proactiva que te ayudan a reducir el riesgo de tiempo de inactividad que podría ser causado por el agotamiento de espacio en disco de tus instancias.
Compute Engine	Costo	Recomendador de descuentos por uso comprometido	El recomendador de descuentos por uso comprometido (CUD Committed used discount, por sus siglas en inglés) te ayuda a optimizar los costos de recursos de los proyectos en tu cuenta de facturación de Google Cloud. Las recomendaciones de CUD se generan automáticamente utilizando una fórmula que analiza métricas de uso históricas y recientes recopiladas por Cloud Billing, e incluye el uso cubierto por compromisos existentes. Puedes aplicar estas recomendaciones para adquirir compromisos adicionales y seguir optimizando los costos de Google Cloud.
Compute Engine	Costo	Recomendador de imagenes personalizadas inactivas	Compute Engine ofrece recomendaciones para ayudarte a identificar recursos como discos persistentes (PDs), direcciones IP e imágenes de disco personalizadas que no se utilizan. Puedes utilizar las recomendaciones de recursos inactivos para minimizar el desperdicio de recursos y reducir tu factura de cómputo. Para los PDs que no se utilizan activamente, puedes crear una copia de seguridad (snapshot) y luego eliminar el recurso. En cuanto a los PDs, imágenes y direcciones IP no utilizadas, puedes eliminarlos si no los necesitas.
Compute Engine	Costo	Discos Persistentes inactivos	Compute Engine ofrece recomendaciones para ayudarte a identificar recursos como discos persistentes (PDs), direcciones IP e imágenes de disco personalizadas que no se utilizan. Puedes utilizar las recomendaciones de recursos inactivos para minimizar el desperdicio de recursos y reducir tu factura de cómputo. Para los PDs que no se utilizan activamente, puedes crear una copia de seguridad (snapshot) y luego eliminar el recurso. En cuanto a los PDs, imágenes y direcciones IP no utilizadas, puedes eliminarlos si no los necesitas.
Compute Engine	Costo	Maquinas Virtuales inactivas	Compute Engine proporciona recomendaciones de máquinas virtuales (VM) inactivas para ayudarte a identificar instancias de VM que no han sido utilizadas durante los últimos 1 a 14 días. Puedes utilizar las recomendaciones de VM inactivas para encontrar y detener las instancias de VM inactivas y así reducir el desperdicio de recursos y disminuir tu factura de cómputo.
Compute Engine	Rendimiento	Recomendador de tipo de máquina para grupos de instancias administradas.	Compute Engine proporciona recomendaciones de tipo de máquina para grupos de instancias administradas (MIGs) para ayudarte a mejorar el rendimiento de las cargas de trabajo y la eficiencia de costos.

Servicio	Categoría	Nombre	Descripción
Compute Engine	Rendimiento	Recomendador de tipo de máquina para VM.	Compute Engine proporciona recomendaciones de tipo de máquina para ayudarte a optimizar la utilización de recursos de tus instancias de máquina virtual (VM). Estas recomendaciones se generan automáticamente en función de las métricas del sistema recopiladas por el servicio de Monitoreo de Cloud durante los últimos 8 días. Utiliza estas recomendaciones para redimensionar el tipo de máquina de tu instancia y así utilizar de manera más eficiente los recursos de la instancia. Esta función también es conocida como recomendaciones de ajuste de tamaño adecuado (rightsizing).
General [1]	Seguridad	Recomendador para notificaciones y gestión de nube	Las recomendaciones de Notificaciones de Asesoramiento supervisan tus Contactos Esenciales y configuraciones de políticas de IAM, y ofrecen recomendaciones basadas en los datos del día anterior. Las recomendaciones incluyen lo siguiente: Si ningún usuario tiene permiso para ver las notificaciones, el recomendador de Notificaciones de Asesoramiento recomienda otorgar acceso a las partes correspondientes dentro de tu organización. Si un principal está listado como un Contacto Esencial de Seguridad pero no tiene permiso para ver las Notificaciones de Asesoramiento en la consola de Google Cloud, el recomendador de Notificaciones de Asesoramiento recomienda otorgar acceso al principal. Las recomendaciones del recomendador de Notificaciones de Asesoramiento no toman en cuenta roles personalizados. Si estás otorgando permisos a un principal para las Notificaciones de Asesoramiento a través de un rol personalizado, ignora o descarta la recomendación.
General	Fiabilidad	Recomendaciones de cambios recientes	Las recomendaciones de cambios recientes identifican de manera automática cambios riesgosos realizados recientemente en recursos en la nube identificados como importantes basados en su uso y otras señales, para ayudar a detectar y mitigar problemas, como interrupciones del servicio, causadas por configuraciones incorrectas de esos recursos importantes en la nube.
General	Mantenibilidad	Recomendador de Descontinuación y Cambios Disruptivos	El recomendador general de descontinuación y cambios disruptivos en la nube te proporciona recomendaciones sobre descontinuaciones y cambios disruptivos en la nube. Identifica los recursos en la nube que se verán afectados por futuras descontinuaciones y cambios disruptivos, al tiempo que proporciona pautas sobre cómo gestionarlos.
General	Mantenibilidad	Recomendador de notificaciones de Error Reporting	El recomendador de Error Reporting busca accidentes recientes en tu proyecto de Google Cloud y proporciona recomendaciones si no has configurado notificaciones de Error Reporting.
General	Mantenibilidad	Recomendador de sugerencias de productos	El recomendador de sugerencias de productos te ayuda a optimizar el uso de tu nube al proporcionarte sugerencias de productos. Esto puede ayudarte a mejorar el rendimiento y la seguridad, y gestionar mejor tus recursos. Basado en las mejores prácticas, analiza el uso actual de productos dentro de cada proyecto y determina cualquier producto adicional que pueda optimizar tu uso.
Google Kubernetes Engine	Costo	Recomendador de clúster de kubernetes inactivos	Puedes identificar los clústeres inactivos estándar de Google Kubernetes Engine (GKE) utilizando las perspectivas y recomendaciones del Recomendador de Clústeres Inactivos. Una vez que verifiques que los clústeres inactivos identificados no están en uso, puedes eliminarlos para ahorrar costos. El Recomendador de Clústeres Inactivos no es relevante para los clústeres de Autopilot, ya que incurren en costos operativos mínimos, ya que solo pagas por los recursos que solicitan tus cargas de trabajo.
Google Kubernetes Engine	Fiabilidad	Recomendador de diagnóstico de GKE.	GKE monitorea tus clústeres y, si existen posibles optimizaciones, proporciona orientación a través de Recommender, un servicio de Google Cloud que genera perspectivas y recomendaciones para el uso de recursos en Google Cloud.

[1] General se refiere a las recomendaciones y/o perspectivas que son generales de Google Cloud y no corresponden a ningún servicio en específico.

Servicio	Categoría	Nombre	Descripción
Google Maps	Mantenibilidad	Recomendador de gestión de proyectos de Google Maps Platform	El recomendador de gestión de proyectos te ayuda a mejorar la salud de tu proyecto de Google Maps Platform.
IAM [2]	Seguridad	Recomendador de permisos excesivos	Las recomendaciones de roles te ayudan a identificar y eliminar permisos excesivos de tus principios, mejorando las configuraciones de seguridad de tus recursos.
Redes	Costo	Recomendador de IPs inactivas	Compute Engine ofrece recomendaciones para ayudarte a identificar recursos como discos persistentes (PDs), direcciones IP e imágenes de disco personalizadas que no se utilizan. Puedes utilizar las recomendaciones de recursos inactivos para minimizar el desperdicio de recursos y reducir tu factura de cómputo. Para los PDs que no se utilizan activamente, puedes crear una copia de seguridad (snapshot) y luego eliminar el recurso. En cuanto a los PDs, imágenes y direcciones IP no utilizadas, puedes eliminarlos si no los necesitas.
Resource manager [3]	Seguridad	Recomendador para proyectos desatendidos	El recomendador de proyectos no supervisados analiza la actividad de uso en los proyectos de tu organización y proporciona recomendaciones que te ayudan a descubrir, recuperar o eliminar proyectos no supervisados.
Resource Manager	Fiabilidad	Recomendador de límites de servicio (cuotas).	El recomendador de límites de servicio analiza el uso de las cuotas de servicio por parte de los proyectos en tu organización y proporciona recomendaciones que te ayudan a identificar recursos que pueden estar cerca de alcanzar sus límites de cuota.
Resource Manager	Fiabilidad	Recomendaciones de riesgo de cambios	Las recomendaciones de riesgo de cambios te ayudan a reducir el riesgo de configuraciones erróneas en la infraestructura en la nube al identificar de manera inteligente cambios riesgosos comunes en tus recursos más importantes y proporcionar recomendaciones para prevenir y mitigar problemas.

Fuente: Elaboración propia y adaptación de (GCP Recommender, 2023)

La tabla **1-2** muestra las perspectivas, también conocidas como *"insights"*, que ofrecen los servicios de Google Cloud. Una perspectiva es una recomendación o información generada a partir del análisis del uso de los recursos de Google Cloud. Estas perspectivas buscan ayudar a los usuarios a optimizar el rendimiento, la seguridad y el costo de sus aplicaciones. Cada perspectiva tiene un *tipo* específico, el cual está relacionado con un único producto y tipo de recurso de Google Cloud. Un único producto puede tener varios tipos de perspectivas, cada una de las cuales proporciona información diferente para un recurso específico. La tabla de perspectivas incluye los siguientes campos: **Servicio:** el servicio de Google Cloud al cual se aplica la perspectiva. **Nombre:** el nombre abreviado de la perspectiva. **Descripción:** una breve explicación de la función de la perspectiva.

Tabla **1-2**.: Conjunto de perspectivas existentes para Google Cloud a la fecha 20 de noviembre de 2024

Servicio	Nombre	Descripción
Cloud Asset	Perspectivas de Cloud Asset	Las Perspectivas de Cloud Asset proporcionan información basada en las políticas de IAM asociadas con los recursos de la organización. Forma parte del servicio de Recommender y se proporciona como el tipo de información `google.cloudasset.asset.Insight`.

[2] Gestión de acceso e identidades
[3] Adminstrador de recursos

Servicio	Nombre	Descripción
Compute Engine	Análisis de recursos inactivos de Compute Engine	Compute Engine proporciona recomendaciones para ayudarte a identificar recursos como discos persistentes (PDs), direcciones IP e imágenes de disco personalizadas que no están siendo utilizadas. Puedes utilizar las recomendaciones de recursos inactivos para minimizar el desperdicio de recursos y reducir tu factura de cómputo. Para los PDs que no están siendo utilizados activamente, puedes crear un snapshot de respaldo y luego eliminar el recurso. En cuanto a los PDs, imágenes y direcciones IP no utilizadas, puedes eliminarlos si no los necesitas, según corresponda.
Compute Engine	Información y recomendaciones sobre grupos de instancias administradas	Las recomendaciones e información sobre grupos de instancias administradas (MIG, por sus siglas en inglés) te ayudan a comprender el uso de CPU y memoria de las instancias de máquinas virtuales (VM) que forman parte de tu MIG. Estas recomendaciones e información se generan automáticamente basadas en métricas del sistema o métricas recopiladas por el servicio de Monitoreo de Cloud.
Compute Engine	Perspectivas de instancias de VM	Las perspectivas de instancias de máquinas virtuales (VM) ayudan a comprender el uso de CPU, memoria y red de tus VM de Compute Engine. Estas perspectivas se generan automáticamente en función de métricas del sistema o métricas recopiladas por Cloud Monitoring.
Dataflow	Perspectivas de Dataflow	Dataflow Insights proporciona información sobre cómo mejorar el rendimiento de los trabajos, reducir costos y solucionar errores. Dataflow Insights forma parte del servicio de Recommender y está disponible a través del tipo `google.dataflow.diagnostics.Insight`.
General	Perspectivas de reporte de errores	El recomendador de Error Reporting genera recomendaciones basadas en información. Puedes obtener esta información utilizando la CLI de Google Cloud o la API de Recommender.
General	Perspectivas de cambios recientes	Las perspectivas de cambios recientes automáticamente señalan cambios riesgosos realizados recientemente en los recursos en la nube identificados como importantes según su uso y otras señales para ayudar a detectar y mitigar problemas, como interrupciones del servicio, causadas por malas configuraciones de esos recursos importantes en la nube.
Google Kubernetes Engine	Perspectivas de diagnóstico de Google Kubernetes Engine	Las perspectivas y recomendaciones sobre obsolescencia están disponibles a través de Recommender, un servicio que ofrece información y recomendaciones para utilizar recursos en Google Cloud.
IAM	Análisis de políticas de IAM	Los análisis de políticas son basadas en aprendizaje de maquina sobre el uso de permisos. Estos análisis pueden ayudar a identificar qué entidades tienen permisos que no necesitan.
IAM	Análisis de cuentas de servicio de IAM	Los análisis de cuentas de servicio son conclusiones sobre qué cuentas de servicio en tu proyecto no han sido utilizadas en los últimos 90 días.
IAM	Movimientos laterales	Las perspectivas sobre movimientos laterales identifican roles que permiten a una cuenta de servicio en un proyecto suplantar a una cuenta de servicio en otro proyecto.
Redes	Perspectivas de reglas de firewall	Firewall Insights ayuda a comprender los patrones de uso de las reglas de firewall. Es posible utilizar esta información para respaldar decisiones sobre la eliminación o modificación de reglas de Firewall para simplificar y asegurar la configuración del firewall.
Redes	Información y recomendaciones del Analizador de Red	El Analizador de Red utiliza comandos de Recomendador. Recomendador es un servicio de Google Cloud que proporciona recomendaciones de uso para productos y servicios de Google Cloud.
Resource Manager	Información sobre proyectos no supervisados	El contenido de una perspectiva de utilización de un proyecto son valores de campo que se utilizan para clasificar la actividad de uso del proyecto y generar recomendaciones de `CLEANUP_PROJECT` y/o `RECLAIM_PROJECT`.

Fuente: Elaboración propia y adaptación de (GCP Recommender, 2023)

Con base en la información proporcionada en las tablas **1-1** y **1-2**, la Tabla **1-3** presenta un resumen de las distintas características que ofrece Google Cloud Recommender para los servicios de Google Cloud. Esta tabla se utilizará en las secciones posteriores de este documento para identificar los productos específicos en los que este trabajo puede contribuir.

Tabla **1-3**.: Conjunto de perspectivas existentes para Google Cloud a la fecha 20 de noviembre de 2024

Servicio	Recomendadores	Perspectivas	Total
BigQuery	2	0	2
Cloud Run	2	0	2
Cloud SQL	6	0	2
Compute Engine	6	3	9
Dataflow	0	1	1
General	5	2	7
Google Kubernetes Engine	2	1	3
Google Maps Platform	1	0	1
IAM	1	3	4
Redes	1	2	3
Resource Manager	3	1	4

Fuente: Elaboración propia y adaptación de (GCP Recommender, 2023)

Las tablas **1-1** y **1-2** muestran el análisis que realiza actualmente Google Cloud Recommender (GCP Recommender, 2023). Con base en este análisis, las siguientes secciones explican el aporte de este trabajo, el cual busca complementar los análisis existentes con nuevas funcionalidades para los servicios de Google Cloud. En particular, se abordarán aquellos servicios que aún no están cubiertos por Google Cloud Recommender (GCP Recommender, 2023) y para los cuales este trabajo puede ofrecer un valor añadido.

1.2. Selección de servicios para el análisis automático de buenas prácticas sobre Google Cloud

En esta sección se detallan los criterios utilizados para la selección de los servicios de Google Cloud que serán objeto de análisis (ver sección 1.2.1). Asimismo, se pone de relieve la importancia que tienen los servicios fundacionales de Google Cloud en la escalabilidad y la configuración adecuada de las organizaciones que operan en esta plataforma. Debido a esta relevancia, se justifica realizar un análisis automatizado de buenas prácticas para estos servicios (ver sección 1.2.2).
Finalmente, tomando como base los criterios establecidos, se seleccionan los servicios que conforman la fundación de Google Cloud y que serán analizados exhaustivamente como parte de este trabajo. Este análisis profundizado permitirá identificar áreas de mejora en la implementación de buenas prácticas para dichos servicios (sección 1.2.3).

1.2.1. Criterios de selección de servicios

Si bien Google Cloud Recommender (GCP Recommender, 2023) ya ofrece análisis de recomendaciones para algunos de sus servicios, como se observa en las tablas **1-1** y **1-2**, la extensa gama de productos de

Google Cloud, que supera los 150 servicios (Google, 2024), cada uno con características y funcionalidades específicas, impide un análisis exhaustivo de todos ellos en el marco de este trabajo. Por lo tanto, se ha optado por delimitar la cantidad de servicios candidatos a ser evaluados y complementados por la solución propuesta. Esta decisión permite no solo optimizar el enfoque del estudio, sino también profundizar en el análisis de aquellos servicios más relevantes para los objetivos planteados.

Para llevar a cabo esta selección de manera estructurada, se ha diseñado un flujo de decisión (figura **1-1**) que permite identificar los servicios más pertinentes. Este flujo se compone de una serie de criterios y tareas específicas que guían el proceso de inclusión o exclusión de los servicios de Google Cloud que serán analizados en este trabajo. A continuación, se explican detalladamente cada una de las fases de este flujo de selección, las cuales permiten filtrar los servicios según su impacto, relevancia en la adopción de buenas prácticas, y su uso dentro de los entornos de nube escalables. Este enfoque sistemático asegura que los servicios seleccionados para el análisis contribuyan de manera significativa a la optimización y mejora de las arquitecturas basadas en Google Cloud.

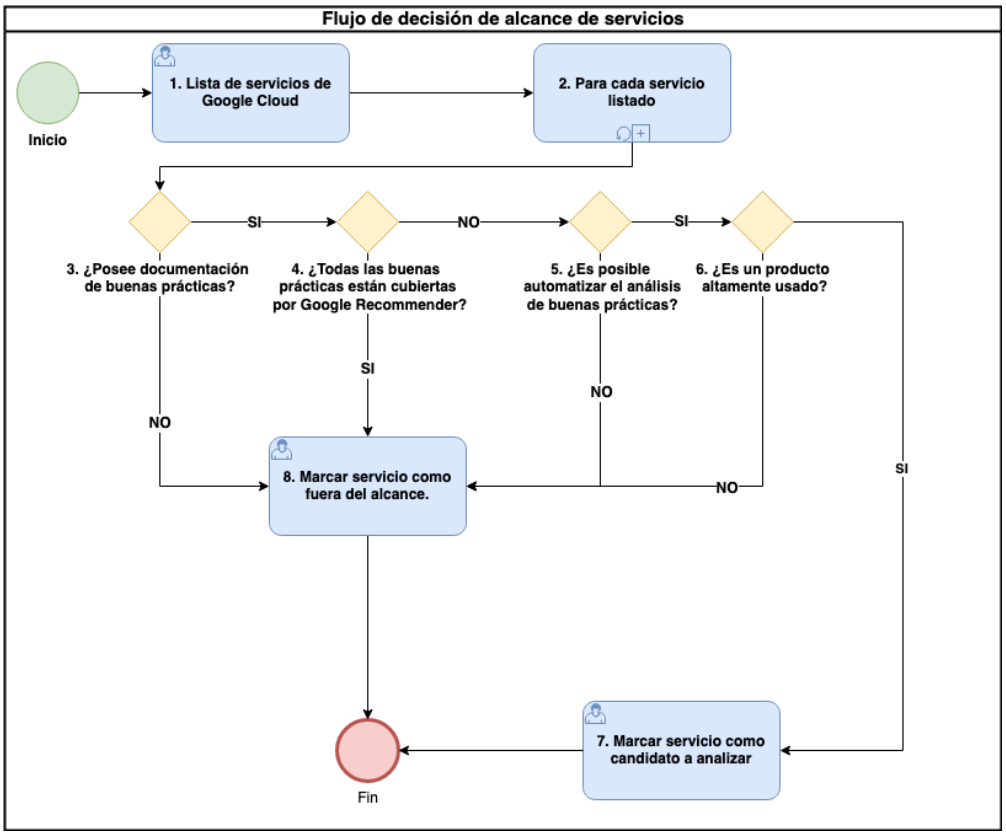


Figura **1-1**.: Diagrama de decisión para selección de servicios de Google cloud.

En la Figura **1-1** se presenta el diagrama de flujo de decisión y en la siguiente lista se explica en detalle cada una de sus tareas:

1. **Listar servicios de Google Cloud**: Se listan todos los servicios de Google Cloud disponibles, utilizando como referencia la documentación oficial de la nube (GCP Service Summary, 2023) .

2. **Listar servicios**: Se listan los servicios que serán evaluados por las siguientes preguntas de flujo 1.

3. **¿Posee documentación de buenas prácticas?**: Se evalúa si el servicio tiene documentación oficial sobre las buenas prácticas de configuración. **SI**: El servicio tiene documentación pública sobre buenas prácticas. Ir a la tarea 4. **NO**: El servicio no tiene documentación pública sobre buenas prácticas. Ir a la tarea 8.

4. **¿Todas las buenas prácticas están cubiertas por Google Recommender?**: Se evalúa si Google Recommender cubre todas las buenas prácticas documentadas. **SI**: El servicio está fuera del alcance porque Google Recommender ya reporta esta configuración. Ir a la tarea 8. **NO**: Google Recommender no cubre todas las buenas prácticas documentadas. Continuar a la pregunta 5.

5. **¿Es posible automatizar el análisis de buenas prácticas?**: Se evalúa si es posible automatizar el análisis de las buenas prácticas en base a la configuración actual del servicio en Google Cloud. **SI**: Automáticamente se puede obtener información del módulo y evaluar su configuración de buenas prácticas con base en la documentación. Continuar con la pregunta 6. **NO**: No es posible obtener información del módulo y/o analizarla automáticamente. Ir a la tarea 8.

6. **¿Es un producto altamente usado?**: Debido al amplio número de servicios disponibles en Google Cloud, es necesario acotar el estudio a un número manejable. La cantidad de usuarios del servicio es un factor clave para seleccionar aquellos con mayor impacto potencial. **SI**: El servicio es candidato a ser analizado. Continuar con la tarea 7. **NO**: El servicio será cubierto en trabajos futuros. Continuar con la tarea 8.

7. **Marcar como servicio candidato a analizar**: Esta tarea consiste en marcar un servicio de Google Cloud como candidato a ser analizado con la herramienta desarrollada en este trabajo de grado.

8. **Marcar módulo como fuera del alcance**: El módulo no será analizado por la herramienta desarrollada en este trabajo de grado. La decisión se toma después de evaluar que el módulo no cumple con los criterios de elegibilidad o no se considera viable su análisis.

1.2.2. Importancia del Cloud Foundations para Google Cloud

Antes de proceder con la selección de los servicios a analizar, es preciso describir el proceso de establecer una infraestructura eficiente y segura en Google Cloud, el cual conlleva una serie de pasos críticos. Inicialmente, se debe configurar la organización, lo que implica establecer el recurso de organización como nivel más alto en la jerarquía de recursos dentro de Google Cloud. Este recurso permite la gestión

centralizada de proyectos, cuentas de facturación y políticas de acceso. Asimismo, se deben crear usuarios administradores con los permisos adecuados y vincular un método de pago para la facturación de los servicios consumidos (figura **1-2**). Dada la importancia crucial de la correcta fundación de los servicios de Google Cloud, este trabajo propone iniciar el análisis precisamente por esta etapa, antes de considerar otros servicios.

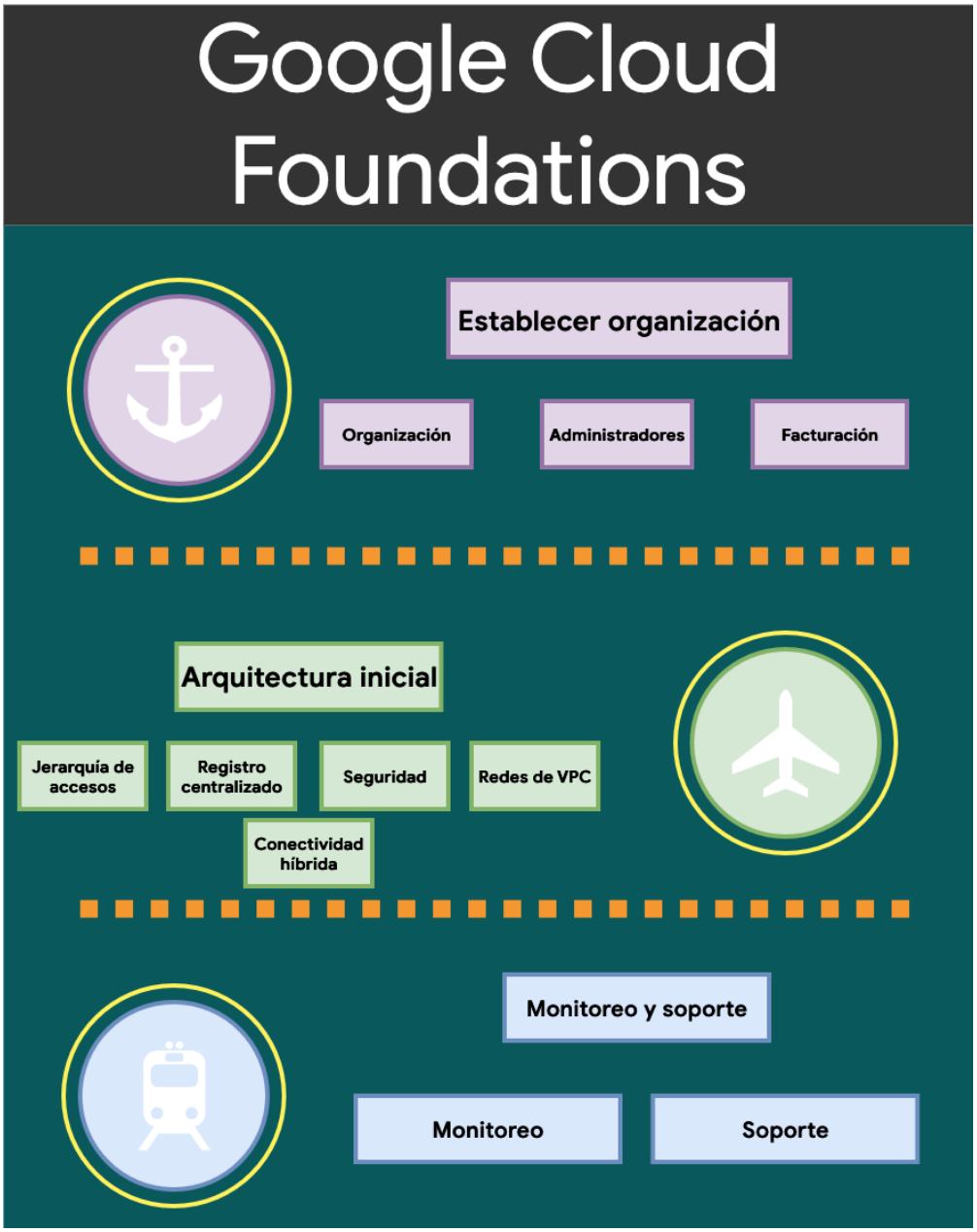


Figura **1-2**.: Módulos de fundación de Google Cloud.

El recurso de organización se crea automáticamente al configurar un servicio de identidad de Google y

vincularlo a su dominio. Existen dos opciones de servicios de identidad: Cloud Identity y Google Workspace. Cloud Identity se encarga de gestionar usuarios y grupos, y permite la federación de identidades entre Google y otros proveedores. Por su parte, Google Workspace ofrece funcionalidades similares, pero incluye herramientas adicionales de productividad y colaboración, como Gmail y Google Drive. Una vez establecida la organización, el siguiente paso es crear una arquitectura inicial. Esto incluye seleccionar una estructura para carpetas y proyectos, asignar permisos de acceso, configurar el registro, aplicar configuraciones de seguridad y establecer la red.

Las carpetas proporcionan un mecanismo para agrupar proyectos y aislarlos entre sí. Pueden representar diferentes departamentos, entornos o equipos dentro de la organización. Los proyectos, por otro lado, contienen los recursos reales de la nube, tales como máquinas virtuales, bases de datos y depósitos de almacenamiento. Google Cloud ofrece varias configuraciones iniciales para la jerarquía de recursos, diseñadas para satisfacer las necesidades de diferentes tipos de organizaciones. Estas configuraciones pueden ir desde pequeñas empresas con entornos centralizados hasta grandes corporaciones con equipos autónomos. Cada configuración incluye una carpeta común destinada a proyectos que contienen recursos compartidos, como el registro y la supervisión.

Después de crear la arquitectura inicial, el siguiente paso es implementar la configuración. Esto se realiza utilizando Terraform, una herramienta de infraestructura como código (IaC, por sus siglas en inglés) que permite definir los recursos en la nube de manera declarativa. Puede implementar la configuración directamente desde la consola de Google Cloud o descargar los archivos de Terraform para personalizarlos e implementarlos dentro de su propio flujo de trabajo. Finalmente, es necesario aplicar las configuraciones de supervisión y soporte. Cloud Monitoring se configura automáticamente para los proyectos en Google Cloud, pero es posible aplicar prácticas recomendadas adicionales para mejorar las capacidades de supervisión. Además, es importante seleccionar un plan de soporte que satisfaga las necesidades de la organización. Google ofrece Basic Support, que es gratuito, y Premium Support, que proporciona opciones de soporte más amplias y completas.

1.2.3. Selección de módulos de fundación para el análisis de recomendaciones

.

En esta sección se presentan los resultados obtenidos tras aplicar el flujo de selección de módulos para el análisis de Google Cloud en los componentes de la fundación. En la figura **1-3**, se destacan en color **Rojo** aquellos módulos que se han excluido del alcance de este trabajo y en color **Azul** los módulos fundacionales de Google Cloud que han sido seleccionados como candidatos para su análisis.

En las subsecciones siguientes se proporcionará una descripción detallada de cada módulo fundacional, junto con la justificación de su inclusión o exclusión en el análisis, abordando los criterios utilizados para esta decisión.

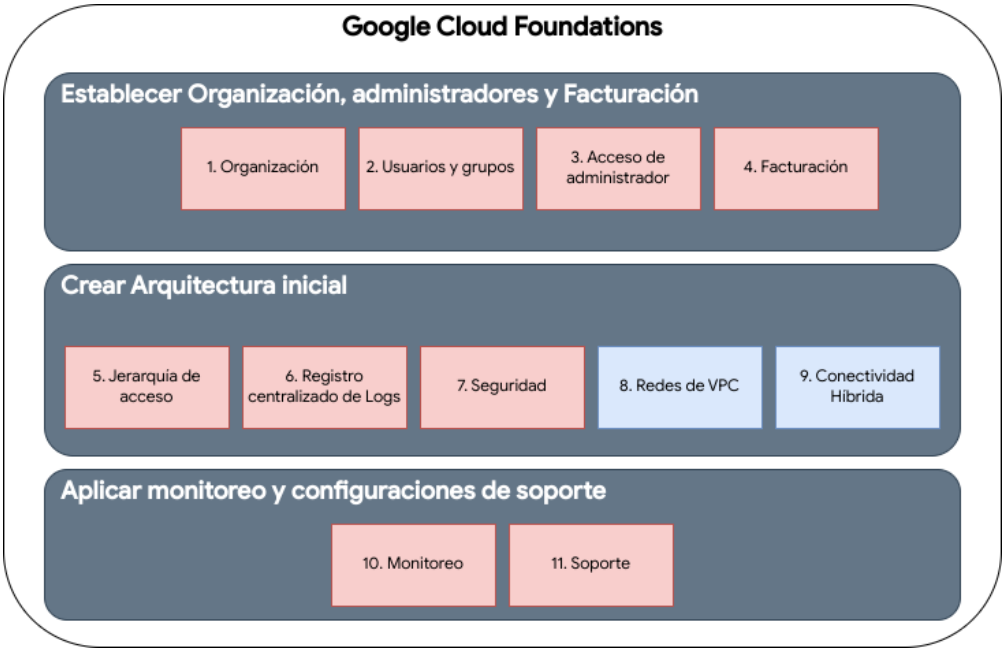


Figura **1-3**.: Diagrama de módulos definidos dentro del alcance del análisis.

Azul: módulos de fundación de Google Cloud candidatos a ser incluidos dentro del alcance del análisis.

Rojo: Modulos de fundación que no se incluyen como parte del alcance del análisis.

1.2.3.1. Fundación de Google Cloud: Redes de VPC

¿Candidato para ser analizado?: Si

Descripción: Una red de Virtual Private Cloud (VPC) es una versión virtual de una red física que se implementa dentro de la red de producción de Google. Una red VPC es un recurso global compuesto por subredes regionales. Las redes VPC proporcionan capacidades de red a los recursos de Google Cloud, tales como instancias de máquinas virtuales de Compute Engine, contenedores de GKE, e instancias del entorno flexible de App Engine.

Shared VPC conecta recursos de varios proyectos a una red VPC común, lo que les permite comunicarse entre sí utilizando las direcciones IP internas de la red. Al usar Shared VPC, se designa un proyecto host y se adjuntan uno o más proyectos de servicio. Las redes de Virtual Private Cloud en el proyecto host se denominan redes de Shared VPC. Es posible usar un proyecto host para gestionar centralmente lo siguiente: Rutas, Reglas de cortafuegos, Conexiones VPN, Subredes. Un proyecto de servicio es cualquier proyecto que esté vinculado a un proyecto host. Es posible compartir subredes, incluidas las de rangos secundarios, entre los proyectos host y de servicio.

Cada red de Shared VPC contiene subredes públicas y privadas: La subred pública puede ser utilizada por instancias orientadas a internet para conectividad externa. La subred privada puede ser utilizada por

instancias internas que no tienen direcciones IP públicas asignadas.

Justificación: El módulo de redes VPC se presenta como un candidato ideal para el análisis de buenas prácticas de infraestructura, dado que los servicios que operan en la nube suelen requerir integración con otros componentes de software ubicados tanto en Google Cloud como en redes on-premises, lo que lo convierte en un producto de alto uso (tarea 6 del flujo **1-1**). Además, debido a su carácter crítico, existe una vasta documentación que establece buenas prácticas específicas para su configuración (tarea 3), y muchas de estas no están completamente cubiertas por el servicio de Google Recommender (GCP Recommender, 2023) (tarea 4).

Otro factor clave que favorece la inclusión de este módulo en el análisis es que las configuraciones actuales de las redes VPC se pueden obtener de manera programática, lo que facilita la evaluación y ajuste de dichas configuraciones en función de las buenas prácticas recomendadas. Por lo tanto, este módulo será parte del análisis y configuración como parte de este trabajo.

1.2.3.2. Fundación de Google Cloud: Conectividad híbrida

¿Candidato para ser analizado?: Si

Descripción: Este proceso crea una VPN de Alta Disponibilidad (HA VPN), que es una solución de alta disponibilidad que se puede implementar rápidamente para transmitir datos a través de internet público.

Este tipo de conexiones permiten establecer enlaces de baja latencia y alta disponibilidad entre redes VPC y redes **on-premises** u otras redes en la nube. Para configurar este tipo de conexiones, se deben establecer los siguientes componentes:

- Puerta de enlace HA VPN de Google Cloud: Un recurso regional que cuenta con dos interfaces, cada una con su propia dirección IP. Es posible especificar el tipo de pila IP, lo que determinará si se admite tráfico IPv6 en la conexión.
- Puerta de enlace VPN del par (peer VPN gateway): La puerta de enlace en la red del par, a la cual se conecta la puerta de enlace HA VPN de Google Cloud. Se deben ingresar las direcciones IP externas que la puerta de enlace del par utiliza para conectarse a Google Cloud.
- Cloud Router: Utiliza el Protocolo de Puerta de Enlace Fronteriza (BGP) para intercambiar rutas dinámicamente entre las redes VPC y las redes del par. Se asigna un Número de Sistema Autónomo (ASN) como identificador para el Cloud Router, y se especifica el ASN que utiliza el enrutador del par.
- Túneles VPN: Conectan la puerta de enlace de Google Cloud con la puerta de enlace del par. Es necesario especificar el protocolo de intercambio de claves de Internet (IKE) que se utilizará para establecer el túnel. Se puede ingresar una clave IKE previamente generada o generar y copiar una nueva.

Justificación: La conectividad de Google Cloud hacia otras nubes o sistemas *on-premises* mediante VPN o Interconnect es un caso de uso bastante común, ya que estos sistemas se comunican de manera segura a través de estos canales. Por ello, Google Cloud ha identificado buenas prácticas para crear estos canales de comunicación híbrida (tarea 3), las cuales no están completamente cubiertas por Google Recommender (tarea 4). Además, es posible obtener la configuración de manera programática (tarea 5) para analizar estas configuraciones, las cuales son ampliamente utilizadas (tarea 6). Por lo tanto, resulta relevante realizar un análisis sobre este módulo fundacional dentro del presente trabajo (ver flujo **1-1**).

La selección de los servicios de VPC y VPN como enfoque principal en este trabajo responde a su papel crítico dentro de la infraestructura de Google Cloud, ya que ambos servicios forman la base de la conectividad, seguridad y comunicación en cualquier implementación en la nube. Las VPC permiten la segmentación y administración de redes de manera escalable y personalizada, mientras que las VPN facilitan la integración segura entre entornos híbridos o multicloud, garantizando la continuidad del negocio y la protección de datos sensibles. Además, su relevancia se extiende a prácticamente todas las arquitecturas en Google Cloud, lo que los convierte en pilares esenciales para evaluar el cumplimiento de buenas prácticas. Al centrar el análisis en estos servicios, el trabajo asegura un impacto significativo al abordar componentes fundamentales que afectan la estabilidad, el rendimiento y la seguridad de las implementaciones en la nube. Para una revisión exhaustiva de todos los servicios de fundación de Google Cloud Platform que no fueron seleccionados en este trabajo, se invita al lector a consultar el anexo B, donde se presenta una lista detallada con información complementaria.

1.3. Requisitos de la aplicación

Los requisitos funcionales 1.3.1 y no funcionales 1.3.2 de este proyecto fueron definidos por el autor, quien cuenta con una sólida experiencia como consultor de Google Cloud especializado en entornos empresariales. Esta experiencia incluye la integración de proyectos reales y el despliegue de soluciones complejas para algunas de las empresas más grandes del mundo. Además, los requisitos definidos se complementaron con los análisis y recomendaciones realizadas por consultores internos de Google Cloud, cuyas observaciones reflejan buenas prácticas y aprendizajes derivados de implementaciones complejas. Sin embargo, estas recomendaciones no se pueden detallar de manera exhaustiva debido a restricciones de confidencialidad asociadas a la organización de Google.

1.3.1. Requisitos funcionales

Los requisitos funcionales son especificaciones esenciales que definen cómo debe comportarse un sistema para cumplir con sus objetivos. Estos requisitos describen las funciones y características que el aplicativo

debe ofrecer al usuario, enfocándose en las tareas específicas que debe realizar y en las interacciones esperadas con el sistema. En el caso de este trabajo, los requisitos funcionales abarcan desde la capacidad del aplicativo para recopilar y analizar configuraciones de infraestructura en Google Cloud, hasta la generación de reportes detallados con recomendaciones basadas en buenas prácticas. Estos requisitos son fundamentales para garantizar que el sistema no solo cumpla su propósito principal, sino que también ofrezca una experiencia de usuario eficiente y alineada con las necesidades del proyecto. A continuación se presenta a alto nivel los requisitos funcionales de la aplicación:

- **Gestión de la configuración de GCP:**
  - Permitir al usuario proporcionar el identificador de la organización de GCP y los proyectos asociados para el análisis.
  - Extraer automáticamente las configuraciones de redes, reglas de firewall, subredes y VPN desde GCP.
  - Validar que los datos de configuración proporcionados sean completos y correctos antes de iniciar el análisis.
- **Análisis de configuraciones:**
  - Identificar configuraciones que no cumplan con las buenas prácticas de GCP.
  - Generar recomendaciones específicas y genéricas basadas en la documentación oficial de Google Cloud.
  - Evaluar configuraciones para redes VPC, VPN, reglas de firewall y subredes.
- **Generación de reportes:**
  - Crear reportes en formato legible para el usuario con un resumen claro de las configuraciones analizadas.
  - Incluir análisis detallados, buenas prácticas, y recomendaciones organizadas por sección.
  - Almacenar los reportes generados en una estructura de carpetas clara y accesible.
  - Proporcionar recomendaciones claras y accionables al usuario para mejorar las configuraciones.
- **Integración con LLMs:**
  - Conectar con modelos de lenguaje para generar análisis de buenas prácticas basados en configuraciones de GCP.
  - Procesar las recomendaciones específicas y genéricas proporcionadas por los LLMs.
- **Personalización del análisis:**

- Permitir la configuración opcional de herramientas adicionales como LangSmith para monitorear la ejecución.
- Solicitar al usuario parámetros clave, como las claves API de OpenAI y otros secretos necesarios.

1.3.2. Requisitos no funcionales

Los requisitos no funcionales son características que determinan cómo debe operar un sistema, abarcando aspectos relacionados con el rendimiento, la usabilidad, la seguridad, la escalabilidad y la mantenibilidad. Estos requisitos complementan a los funcionales al establecer estándares de calidad y restricciones técnicas que aseguran un funcionamiento eficiente y confiable del sistema. En el contexto de este proyecto, los requisitos no funcionales incluyen garantizar tiempos de respuesta adecuados durante el análisis de configuraciones, proteger los datos sensibles mediante medidas de seguridad avanzadas, y asegurar que el sistema sea escalable para manejar infraestructuras de mayor complejidad en el futuro. Esta lista se presenta a continuación:

- El sistema debe poder analizar organizaciones con múltiples proyectos y configuraciones complejas sin degradar el rendimiento.
- Completar el análisis y generación de reportes en un tiempo razonable, idealmente menos de 15 minutos para organizaciones medianas.
- La interfaz debe ser clara y accesible para usuarios con conocimientos básicos de GCP.
- El aplicativo debe ser ejecutable en diferentes entornos, como sistemas operativos Windows, macOS y Linux.
- El código debe estar bien documentado y estructurado para facilitar futuras mejoras y expansiones.
- Las dependencias externas deben estar actualizadas y gestionadas correctamente.
- Proteger los datos sensibles proporcionados por el usuario, como claves API y configuraciones de red.
- El sistema debe ser compatible con las versiones actuales de las APIs de GCP.
- Registrar la ejecución del análisis para facilitar auditorías y seguimiento del comportamiento del sistema.

Tras finalizar el análisis de los módulos de fundación de Google Cloud Platform considerados para este trabajo y la definición de los requisitos funcionales y no funcionales, se procederá a describir la herramienta de análisis en los capítulos siguientes. Esta herramienta se aplicará a los módulos seleccionados en esta sección, cuya elección se basó en el flujo de selección detallado en la sección 1.2.1.

1.4. Justificación del uso de LLMs para el análisis de buenas prácticas

Para diseñar una solución que analice configuraciones de infraestructura en Google Cloud y ofrezca recomendaciones basadas en mejores prácticas, es crucial seleccionar una metodología que maximice la flexibilidad y capacidad de adaptación. Aunque las soluciones basadas en reglas son ampliamente utilizadas en la industria, este trabajo opta por el uso de Modelos de Lenguaje Extensos debido a las ventajas sustanciales que ofrecen en comparación con enfoques tradicionales.

El enfoque basado en reglas se construye mediante la codificación explícita de criterios y patrones en un conjunto de instrucciones programáticas. Herramientas como Config Validator Google (Google Config Validator, 2024), Terraform Validator (Google Terraform Validator, 2024), o Policy Analyzer (Google Policy Analyzer, 2024) permiten evaluar configuraciones contra reglas predefinidas. Aunque este enfoque tiene ventajas en términos de simplicidad y facilidad de interpretación, presenta algunas limitaciones:

- La creación de reglas específicas para cada escenario requiere un esfuerzo significativo, especialmente en infraestructuras complejas y cambiantes. Las reglas deben actualizarse continuamente para reflejar cambios en los servicios, productos o regulaciones.
- En un entorno como Google Cloud, donde los servicios evolucionan rápidamente, mantener una base de reglas al día puede convertirse en una tarea titánica.
- Las herramientas basadas en reglas son rígidas. Si surge un caso que no encaja perfectamente en las reglas predefinidas, la herramienta puede fallar en proporcionar recomendaciones útiles.
- No pueden inferir relaciones complejas ni generar recomendaciones que combinen contexto, conocimiento implícito y creatividad, aspectos que son clave en configuraciones empresariales personalizadas.
- Las herramientas basadas en reglas evalúan configuraciones de forma binaria: cumplen o no cumplen. No ofrecen un análisis más profundo o explicaciones detalladas que podrían mejorar la comprensión del usuario sobre las mejores prácticas.
- Config Validator (Google Config Validator, 2024) verifica si las configuraciones cumplen con políticas, pero no ofrece sugerencias proactivas o alternativas más optimizadas.

Los LLMs como GPT-4 representan una nueva generación de herramientas con capacidades avanzadas de procesamiento de lenguaje natural (Chang, 2023). Estas capacidades los convierten en una opción ideal para abordar la complejidad y diversidad de las configuraciones en Google Cloud. Algunas de estas capacidades se mencionan a continuación (Chang, 2023):

- Los LLMs pueden analizar configuraciones extensas y detectar patrones o relaciones implícitas entre diferentes elementos de la infraestructura.

- A diferencia de las reglas predefinidas, los LLMs pueden ajustarse dinámicamente a diferentes escenarios mediante entrenamiento adicional o ajuste fino (fine-tuning)
- Los LLMs no solo identifican problemas; también proporcionan recomendaciones detalladas, con explicaciones y alternativas fundamentadas. Si una regla de firewall está demasiado abierta, un LLM puede sugerir configuraciones alternativas más seguras.
- Los LLMs no solo identifican problemas; también proporcionan recomendaciones detalladas, con explicaciones y alternativas fundamentadas.
- Los LLMs pueden incorporar conocimientos adicionales provenientes de documentación oficial, artículos técnicos o incluso experiencias previas incluidas en su entrenamiento.
- Los LLMs permiten un enfoque conversacional, lo que facilita la interacción con usuarios menos técnicos y promueve un entendimiento más profundo de las recomendaciones.

La herramienta diseñada en este trabajo se basa en un flujo lógico que combina las capacidades de un modelo de lenguaje con un proceso de análisis automatizado, el cual se detalla en el capítulo 2, incluyendo el algoritmo empleado y la estrategia para realizar el análisis de buenas prácticas mediante LLMs.
La elección de un enfoque basado en LLMs se sustenta en un análisis comparativo con herramientas tradicionales basadas en reglas, como Config Validator (Google Config Validator, 2024) o Terraform Validator (Google Terraform Validator, 2024). Mientras que estas últimas presentan limitaciones al lidiar con configuraciones complejas, los LLMs ofrecen una solución más robusta, adaptable y rica en recomendaciones. Esta capacidad no solo mejora la experiencia del usuario al proporcionar análisis más profundos, sino que también permite a la herramienta adaptarse a la evolución constante de los entornos en la nube, manteniéndose relevante frente a las demandas actuales y futuras.

2. Diseño e implementación de herramienta de análisis automático de buenas prácticas en Google Cloud

En este capítulo se detalla el diseño a alto nivel de la herramienta para el análisis automático de buenas prácticas en Google Cloud. La estrategia y estructura del diseño se ilustran en el diagrama de la figura **2-1**.

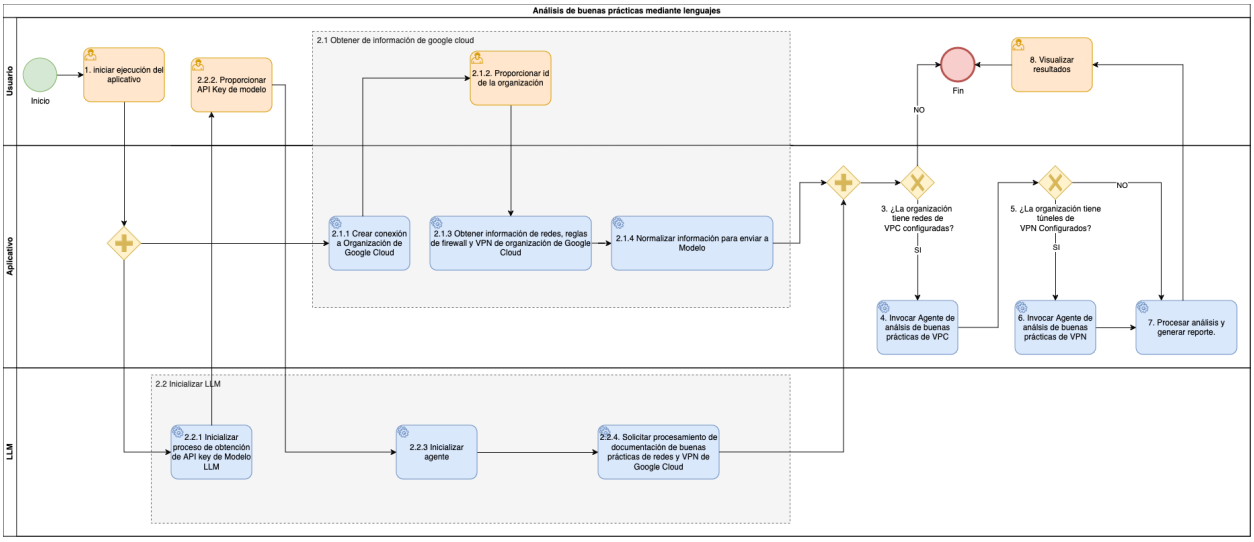


Figura **2-1**.: Estrategia de alto nivel de la herramienta para el análisis de buenas prácticas en Google Cloud

La figura **2-1** muestra, a través de un diagrama de flujo de negocio, las tareas que componen el funcionamiento de la herramienta. Estas tareas representan los pasos esenciales que realiza el sistema para evaluar las configuraciones de Google Cloud y su conformidad con las buenas prácticas recomendadas. La arquitectura sigue una serie de fases, desde la recolección de datos de configuración hasta la generación de reportes, proporcionando una visión general del estado de cada módulo.

En la sección 2.1 se describe en profundidad la responsabilidad de cada tarea que compone el flujo de la herramienta, detallando las acciones específicas que se llevan a cabo en cada etapa del proceso. Esta sección permite comprender mejor el flujo de operaciones diseñado para alcanzar los objetivos de análisis. Posteriormente, en la sección 2.2, se profundiza en la implementación del código fuente de la herramienta.

Esta implementación fue diseñada para seguir el flujo de operaciones previamente planteado, asegurando que cada paso del análisis esté alineado con los objetivos de identificación y recomendación de buenas prácticas en Google Cloud.
La sección 2.2 ofrece un análisis detallado de la implementación realizada para el diseño presentado en esta sección 2.1. Se ha decidido separar el detalle técnico y las herramientas utilizadas del planteamiento estratégico y el diseño conceptual para permitir una comprensión clara del objetivo del proyecto, sin necesidad de adentrarse en los detalles específicos de las herramientas y técnicas empleadas. Esta división facilita la comprensión general del enfoque del proyecto antes de profundizar en los aspectos técnicos.

2.1. Diseño e implementación de la Herramienta de Análisis

Esta sección se dedica a explicar, a alto nivel, cada una de las tareas que componen el flujo de negocio presentado en la figura **2-1**. El objetivo es comprender el propósito semántico de cada tarea y, en consecuencia, el aporte que se busca con este trabajo. Posteriormente, en la sección 2.2, se profundizará en la implementación del diseño propuesto.

2.1.1. Actores del flujo

La figura **III-1** muestra tres divisiones horizontales que representan los actores principales involucrados en el proceso de la herramienta. Estas divisiones contienen las tareas que corresponden a cada uno de los siguientes actores:

- **Usuario**: representa al usuario final a quien se presentarán los resultados del análisis realizado por la herramienta.
- **Aplicativo**: es el software o herramienta implementada que coordina el flujo de ejecución y orquesta las tareas necesarias para completar el análisis.
- **LLM (Gran Modelo de Lenguaje)**: aunque forma parte del aplicativo y es invocado por éste, se ha destacado como un actor independiente para diferenciar aquellas tareas que requieren la inicialización y uso específico del modelo de lenguaje.

La inclusión de un Gran Modelo de Lenguaje (LLM) permite simplificar el proceso de extracción de buenas prácticas desde la documentación. Realizar manualmente la síntesis, interpretación y extracción de estas buenas prácticas podría ser un proceso largo y tedioso para el usuario final. Además, dado que la documentación se actualiza frecuentemente, el LLM facilita el proceso de análisis al permitir al usuario ahorrar tiempo y esfuerzo en la revisión y lectura continua de la documentación.

2.1.2. Iniciar ejecución del aplicativo

Tarea ejecutada por el actor: **Usuario**.
Esta tarea representa el inicio de la ejecución de la herramienta. El usuario, quien debe tener una cuenta en Google Cloud, lanza manualmente el aplicativo desde un computador. Dado que la ejecución no se ha configurado para iniciarse automáticamente, es el usuario quien determina cuándo y con qué frecuencia desea realizar el análisis de buenas prácticas de Google Cloud, adaptándose a sus necesidades o requerimientos específicos.

2.1.3. Obtener información de Google Cloud

Tarea ejecutada por los actores: **Aplicativo** y **Usuario**.
El flujo ilustrado en la figura **2-1** plantea esta tarea como un subproceso que agrupa diversas subtareas orientadas a la obtención de información desde Google Cloud. En este diseño, se propone que la aplicación se conecte de manera programática a la organización de Google Cloud del usuario para llevar a cabo el análisis de buenas prácticas. Las tareas descritas a continuación corresponden a este proceso de conexión y recopilación de información desde Google Cloud, permitiendo así una evaluación integral y automatizada de la infraestructura del usuario.

2.1.3.1. Crear conexión a organización de Google Cloud

Subtarea ejecutada por el actor: **Aplicativo**.
El objetivo de esta subtarea es establecer la conexión con Google Cloud, iniciando de manera programática un proceso de acceso al API de Google Cloud. Este proceso permite adquirir los permisos necesarios para recopilar la información que posteriormente será analizada por el aplicativo.

2.1.3.2. Proporcionar id de la organización

Subtarea Ejecutada por el actor: **Usuario**.
El identificador de la organización es un número de 12 dígitos que no inicia con 0 y que permite identificar que los recursos están asociados a una organización en Google Cloud. Usualmente, Google Cloud asocia a las organizaciones mediante la validación de un dominio DNS público. Por ejemplo, si una empresa *X* posee el dominio *www.empresax.com*, esta deberá validar la propiedad del dominio ante Google Cloud. Una vez completada esta verificación, Google Cloud asignará un identificador único de organización a dicho dominio.
Es esencial que el usuario proporcione este identificador, ya que es mediante él que el aplicativo podrá obtener, de manera recursiva, todos los proyectos e información de redes asociados, que serán analizados en la plataforma.

2.1.3.3. Obtener información de redes, reglas de firewall y VPN

Subtarea ejecutada por el actor: **Aplicativo**.

Una vez que la conexión a Google Cloud esté establecida y se haya proporcionado el identificador de la organización, el aplicativo ejecutará las siguientes tareas de forma secuencial:

- **Obtener proyectos de forma recursiva:** Los proyectos de Google Cloud contienen los recursos y servicios proporcionados por la plataforma. Por lo tanto, es esencial obtener los identificadores de los proyectos de manera recursiva, utilizando el identificador de organización proporcionado en la subtarea anterior.
- **Obtener información de redes de VPC [1]:** Para cada proyecto listado de manera recursiva, se obtienen las redes de VPC, las subredes y las reglas de firewall configuradas en cada una de ellas.
- **Obtener información de conectividad híbrida de VPN:** Una vez identificados los proyectos que contienen redes de VPC, que son indispensables para crear una conexión VPN, el aplicativo llama al API de Google Cloud de VPN en esos proyectos para obtener información sobre los túneles de VPN creados en dichas redes.

2.1.3.4. Normalizar información para enviar a Modelo

Tarea ejecutada por el actor: **Aplicativo**.
En esta tarea, la información obtenida por el aplicativo se procesa y se estructura en un formato adecuado para ser enviado al modelo. Esto se debe a que el API de Google Cloud devuelve los datos de las configuraciones en un formato específico que requiere normalización y ajuste antes de su envío. Este proceso de transformación asegura que los datos estén organizados y listos para ser analizados de manera efectiva por el modelo.

2.1.4. Inicializar LLM

Tarea ejecutada por los actores: **LLM** y **Usuario**.
El flujo mostrado en la figura **2-1** define esta tarea como un subproceso que agrupa una serie de subtareas dedicadas a la inicialización del gran modelo de lenguaje y su correspondiente agente. Esto se debe a que la herramienta destinada al análisis de la información extraída automáticamente de Google Cloud requiere procesar grandes volúmenes de datos, como la documentación de buenas prácticas y las configuraciones detalladas de red. Para gestionar esta complejidad, el modelo de lenguaje realizará esta tarea de análisis.

[1] VPC es el acrónimo en inglés para ”Virtual Private Cloud,”que representa la creación de una nube privada mediante la configuración de red.

Este enfoque permite diseñar un sistema adaptable, fundamentado en las capacidades de los modelos de lenguaje avanzados, en lugar de depender de un sistema estático basado en reglas que necesitaría ajustes frecuentes cada vez que la documentación cambia. Las subtareas de este subproceso, descritas a continuación, tienen como objetivo inicializar y configurar el gran modelo de lenguaje para cumplir eficazmente con el propósito de analizar las configuraciones, que constituye el objetivo principal de este trabajo.

2.1.4.1. Inicializar modelo de lenguaje mediante API Key

Tarea ejecutada por el actor: **LLM**.
Los grandes modelos de lenguaje demandan un poder computacional considerable y son desarrollados y mantenidos por gigantes tecnológicos como Google, Microsoft y Meta, entre otros. Para acceder a sus capacidades, es necesario contar con un API Key que permite a estas compañías monetizar el uso de sus modelos, así como rastrear y monitorear los llamados y otras actividades. Además, el API Key identifica qué usuario y/o aplicación está accediendo a los modelos. Aunque varios de estos modelos ofrecen una versión gratuita, igualmente requieren un API Key para su utilización programática, es decir, cuando son empleados por aplicaciones distintas a sus interfaces gráficas (si estas existen).
Por lo tanto, es fundamental que el modelo de lenguaje sea inicializado con el correspondiente API Key. En esta tarea, el aplicativo invoca el proceso de inicialización del modelo mediante un API Key, y a continuación permite al usuario ingresar la clave de API requerida para habilitar su funcionamiento.

2.1.4.2. Proporcionar API Key de modelo

Tarea ejecutada por el actor: **Usuario**.
Como se resaltó en la tarea anterior, contar con un API Key es indispensable para poder aprovechar las capacidades de los grandes modelos de lenguaje. Los API Key son datos sensibles cuya pérdida puede resultar en consecuencias financieras y en riesgos de seguridad, como la suplantación de identidad; por esta razón, su custodia y uso deben gestionarse con sumo cuidado. Para mejorar la seguridad, se delega al usuario la responsabilidad de proporcionar el API Key durante la ejecución del programa, evitando así que estos datos sensibles se almacenen de manera insegura en el código fuente.
Alternativamente, el API Key podría obtenerse de un sistema de administración de secretos al que la aplicación tenga acceso. Un ejemplo adecuado sería el administrador de secretos de Google Cloud, el cual proporciona una manera segura de almacenar y gestionar credenciales y claves de acceso, minimizando los riesgos asociados a la exposición de estos datos.

2.1.4.3. Inicializar agente

Tarea ejecutada por el actor: **LLM**.
Los agentes, como se explica en la sección IV.4, permiten controlar el flujo de ejecución, el contexto y las

tareas que el modelo de lenguaje procesará. En esta tarea, se inicializa el agente que se encargará de ejecutar el análisis de buenas prácticas. Este agente, durante el análisis, irá tomando decisiones sobre qué acciones realizar: ya sea utilizar el modelo de lenguaje para interpretar la información de la configuración de la red, o ejecutar tareas programáticas para, por ejemplo, acceder a la API de Google Cloud y obtener datos, comparar valores o generar reportes. Este proceso de inicialización es crucial, ya que establece el marco de ejecución para el análisis, permitiendo que el agente coordine de forma eficiente las diferentes etapas del proceso.

2.1.4.4. Solicitar procesamiento de documentación de buenas prácticas de redes y VPN de Google Cloud

Tarea ejecutada por el actor: **LLM**.
El objetivo de esta tarea es que el agente consulte la documentación en línea sobre las mejores prácticas para redes de VPC (GCP VPN Best Practices, 2024) y VPN (GCP VPN Best Practices, 2024) de Google Cloud. La finalidad es extraer su texto y agregarlo al contexto de la conversación con el modelo de lenguaje de gran escala. Esto permitirá al modelo tomar en cuenta esta documentación específica durante el análisis. Es fundamental obtener esta información en tiempo real para que, en caso de que el equipo de Google Cloud actualice la documentación, el contexto proporcionado al modelo refleje siempre los contenidos más recientes, optimizando así la precisión y actualidad del análisis.

2.1.5. ¿La Organización tiene redes de VPC configuradas?

Tarea ejecutada por el actor: **Aplicativo**.
Este paso condicional del flujo en el aplicativo permite determinar si existen redes de VPC configuradas en la organización del usuario, una vez completado el subproceso descrito en 2.1.3. Si ningún proyecto dentro de la organización del usuario contiene subredes configuradas, implica que no posee recursos que requieran de una VPC o que los servicios de Google Cloud en uso no dependen de la creación de redes virtuales. En el caso de que se detecten redes de VPC configuradas, el sistema procederá a invocar al agente de buenas prácticas de VPC en la tarea 2.1.6. De lo contrario, el aplicativo finalizará la ejecución, ya que no habrá configuraciones de red para analizar en la cuenta del usuario. Además, no se podrá iniciar el análisis de buenas prácticas para VPN, dado que las VPN requieren una configuración previa de redes de VPC.
Es fundamental que el modelo de lenguaje de gran escala (LLM) sea invocado solo cuando realmente se necesita, ya que el uso excesivo del modelo puede generar costos considerables debido al consumo del API Key.

2.1.6. Invocar Agente de análsis de buenas prácticas de VPC

Tarea ejecutada por el actor: **Aplicativo**.
Una vez ejecutados los dos subprocesos de inicialización —es decir, la obtención de la información de Google Cloud (2.1.3) y la inicialización del agente de LLM (2.1.4)— el aplicativo solicitará al agente que realice un análisis de la configuración actual de la red de VPN en la organización. Este análisis se basará tanto en la documentación de buenas prácticas de VPC como en los datos específicos de la configuración obtenida previamente. El objetivo de esta solicitud es que el agente genere un concepto detallado sobre el estado de la configuración, el cual se almacenará temporalmente en el aplicativo para, en una fase posterior, integrarlo en un reporte final que será presentado al usuario final.

2.1.7. ¿La Organización tiene túneles de VPN configurados?

Tarea ejecutada por el actor: **Aplicativo**.
Este condicional se evalúa una vez se ha confirmado que el proyecto en análisis cuenta con redes de VPC configuradas; de lo contrario, la ejecución del aplicativo ya habría concluido. En esta fase, la aplicación recorrerá todos los proyectos con redes de VPC en busca de túneles de VPN configurados. Si no encuentra ningún túnel de VPN, la aplicación concluirá el análisis de redes; de hallarse túneles de VPN, la aplicación procederá a invocar la tarea correspondiente para realizar el análisis de buenas prácticas de VPN a través del agente, detallado en la sección 2.1.8.

2.1.8. Invocar Agente de análsis de buenas prácticas de VPN

Tarea ejecutada por el actor: **Aplicativo**.
Esta tarea se encarga de invocar el análisis de buenas prácticas de VPN basado en la configuración que obtuvo directamente de Google Cloud, en esta el agente invoca al modelo de lenguaje para que basado en la documentación y basado en la configuración existente, emita un concepto con respecto a la configuración actual, el resultado del texto emitido por el agente será almacenado por el aplicativo de forma temporal, para ser presentado en el reporte final construído posteriormente en la tarea 2.1.9

2.1.9. Procesar análisis y generar reporte

Tarea ejecutada por el actor: **Aplicativo**.
Hasta este punto, todos los pasos previos han sido ejecutados automáticamente, invocando tanto el API de Google Cloud como el agente y el modelo de lenguaje. Sin embargo, para facilitar una visualización comprensible y útil para el usuario, esta tarea final genera un reporte legible, que permite al usuario revisar la información y tomar decisiones informadas.

El reporte debe incluir un resumen de las redes de VPC identificadas, los túneles de VPN encontrados, así como los conceptos y recomendaciones emitidos por el gran modelo de lenguaje. En caso de que la organización no cuente con redes de VPC y el análisis no haya sido posible, el aplicativo debe generar un reporte que indique que, en el momento, no se encontró ninguna configuración en la organización para ser analizada.

2.1.10. Visualizar resultados

Tarea ejecutada por el actor: **Usuario**.
Esta tarea consiste únicamente en que el usuario abra el reporte generado durante la tarea 2.1.9. En esta última etapa, el sistema ya ha recopilado, procesado y presentado toda la información relevante en un formato de reporte visual, accesible y estructurado. Al abrir el reporte, el usuario podrá revisar los resultados del análisis de buenas prácticas en Google Cloud, incluyendo detalles de redes de VPC, configuraciones de VPN, recomendaciones emitidas por el modelo LLM y cualquier otro hallazgo relevante. Esta acción facilita la revisión y toma de decisiones basadas en el análisis detallado de la infraestructura y prácticas de Google Cloud, con un enfoque visual y práctico para el usuario final.

2.2. Implementación de herramienta de análisis de buenas prácticas de Google Cloud

En esta sección se presenta a profundidad el detalle técnico de la implementación de la herramienta de anásis de buenas prácticas de Google Cloud.

2.2.1. Tecnologías usadas para la implementación

2.2.1.1. Uso de Python y Google Cloud SDK

Python fue elegido para la extracción de información de Google Cloud debido a su flexibilidad, simplicidad y extensa biblioteca de herramientas especializadas en la integración con servicios en la nube. Python es un lenguaje ampliamente utilizado en la industria de la tecnología y, en particular, en proyectos de ciencia de datos y desarrollo de aplicaciones en la nube, donde destaca por su capacidad de manipulación de datos y la integración de APIs. Google ofrece un SDK específico para Python, conocido como Google Cloud SDK, el cual incluye una serie de bibliotecas diseñadas específicamente para facilitar la conexión y la interacción con los diversos servicios de Google Cloud.
El Google Cloud SDK para Python proporciona acceso nativo y seguro a los recursos de Google Cloud, permitiendo que la herramienta implemente un flujo programático de extracción de datos de manera eficiente.

Este SDK se encarga de gestionar los permisos y la autenticación, simplificando el acceso a la infraestructura de la nube mediante el uso de credenciales preconfiguradas que permiten una integración segura con la plataforma de Google Cloud. Además, su capacidad para realizar llamadas a las APIs de los servicios de red y VPN es fundamental para el objetivo de este proyecto, que se centra en analizar configuraciones de red en busca de buenas prácticas. La capacidad de Python para manejar solicitudes HTTP de manera asíncrona, junto con el soporte nativo del SDK para autenticación y manejo de errores, hace de este entorno una elección ideal para realizar consultas programáticas y recuperar datos de manera eficiente.

2.2.1.2. Uso de LangGraph y LangChain para la Inicialización del Agente

LangGraph y LangChain fueron seleccionados para la inicialización y gestión del agente que orquestará el flujo de análisis debido a sus capacidades avanzadas de gestión de contextos y flujos de trabajo en aplicaciones de inteligencia artificial. LangChain permite la implementación de agentes en aplicaciones que requieren modelos de lenguaje de gran escala (LLMs) para ejecutar tareas complejas de manera contextual. LangGraph se complementa en este aspecto al proporcionar una estructura de grafos de flujo que puede representar las decisiones y dependencias que deben seguirse al invocar un modelo de lenguaje en un entorno complejo y condicionado, como lo es la configuración de redes de Google Cloud.
El agente inicializado en LangChain se utiliza para automatizar el proceso de invocación y consulta al modelo de lenguaje, manteniendo el contexto de la tarea de análisis y permitiendo una secuencia lógica de decisiones que el agente puede tomar en tiempo real. Esta característica es esencial para el proyecto, ya que LangChain proporciona un marco de trabajo modular donde se puede definir un flujo de decisiones y condiciones que guíen al modelo de lenguaje en la generación de recomendaciones de configuración de red. La posibilidad de utilizar LangGraph y LangChain en conjunto permite, además, definir los elementos específicos del flujo (como los condicionales y subtareas) de manera explícita, brindando una transparencia que contribuye a un análisis repetible y verificable. Así, la elección de estas herramientas es clave para garantizar una administración coherente y lógica de las interacciones entre el modelo de lenguaje y la infraestructura subyacente de Google Cloud.

2.2.1.3. Uso del Modelo de Lenguaje OpenAI ChatGPT-4

El modelo de lenguaje ChatGPT-4 de OpenAI (Open Ai, 2024) fue seleccionado debido a su capacidad para analizar contextos complejos y generar recomendaciones basadas en la comprensión profunda de textos técnicos y documentación de buenas prácticas. ChatGPT-4 ha demostrado tener una notable precisión en el procesamiento de lenguaje natural y, en particular, en la interpretación de preguntas técnicas y en la generación de respuestas con un nivel de detalle adecuado. La elección de ChatGPT-4 (Open Ai, 2024) está justificada por su entrenamiento en un corpus de datos amplio y diverso, lo que le permite abordar consultas sobre configuraciones de redes y seguridad de una manera generalizada y efectiva.

Una de las razones por las que ChatGPT-4 es ideal para este proyecto es su capacidad para generar resúmenes, extraer puntos clave y hacer análisis de configuraciones complejas. Para este caso de uso, donde se analizan y contrastan configuraciones de Google Cloud en relación con las mejores prácticas de redes y VPN, el modelo proporciona una interpretación confiable y adaptable. Gracias a su diseño y entrenamiento, ChatGPT-4 también es eficaz en la generación de análisis y recomendaciones de configuración, lo que lo hace particularmente útil para organizaciones que buscan evaluar configuraciones sin necesidad de un análisis completamente manual (Open Ai, 2024).
Además, ChatGPT-4 está respaldado por una API robusta que permite integrar su funcionalidad dentro de aplicaciones de forma segura y escalable. En este sentido, su uso dentro de este proyecto permite que el sistema aproveche el procesamiento de lenguaje de una automática y controlada, al mismo tiempo que proporciona flexibilidad para adaptarse a diferentes necesidades de análisis. Este modelo se convierte en una herramienta esencial para garantizar que el usuario reciba un análisis que no solo es informativo sino también comprensible, lo que contribuye significativamente al propósito de la herramienta desarrollada en este proyecto (Open Ai, 2024).

2.2.2. Implementación

En esta sección se presentan los módulos de Python desarrollados para implementar de automática la estrategia descrita en la sección 2.1. Cada módulo incluye su respectiva imagen, ilustrando los métodos y atributos de cada clase, así como un texto explicativo sobre las funciones y responsabilidades de la clase. En caso necesario, se hace referencia a los pasos de la estrategia de alto nivel establecidos en la sección 2.1. Es importante señalar que, para mantener el contenido resumido, no se muestra el código fuente de cada clase en detalle. Sin embargo, el código fuente completo puede ser consultado en el repositorio público en GitHub, donde está disponible la implementación detallada de cada módulo: `https://github.com/roger8849/gcp_infra_best_practices_analyzer/tree/main/gcp_configurations_analyzer`.

2.2.2.1. Modulo prinicipal App

Este módulo contiene la clase `App` (ver figura **2-2**), que funciona como la clase principal encargada de iniciar la ejecución del aplicativo. En particular, el método `main` de esta clase activa los métodos de inicialización tanto del aplicativo como del agente de lenguaje. Por lo tanto, esta clase implementa la tarea descrita en la sección 2.1.2, permitiendo la configuración y el despliegue inicial de los componentes necesarios para el análisis de buenas prácticas de Google Cloud.

2.2.2.2. Modelos

El módulo de modelos (ver figura **2-3**) incluye las siguientes dos clases:

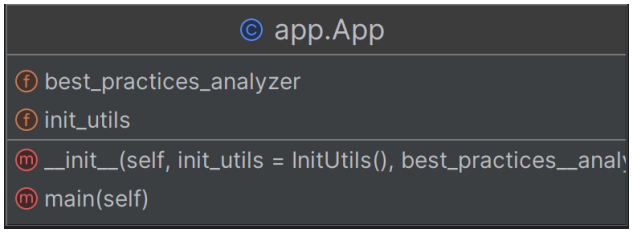


Figura **2-2**.: Clase principal App

- `GCPBestPracticesState`: Esta clase se encarga de almacenar en memoria la información recopilada por el aplicativo a lo largo de su ejecución, incluyendo tanto los datos extraídos de Google Cloud como los proporcionados por el agente de LLM. En esta clase se almacenan los datos de los proyectos dentro de la organización, los proyectos que contienen redes, las redes de VPC, las reglas de firewall, las subredes, los túneles de VPN, las mejores prácticas de VPC y VPN, la información de Google Cloud ya normalizada, y las recomendaciones que el LLM genera para la configuración tanto de VPC como de VPN, además del reporte final.
- `ApplicationConfiguration`: Esta clase guarda en memoria la configuración requerida para ejecutar el aplicativo. Contiene, entre otros elementos, la API Key del modelo de lenguaje LLM, la lista de regiones de Google Cloud disponibles para la organización, el identificador del proyecto principal, el modelo de lenguaje inicializado, y el identificador de la organización.

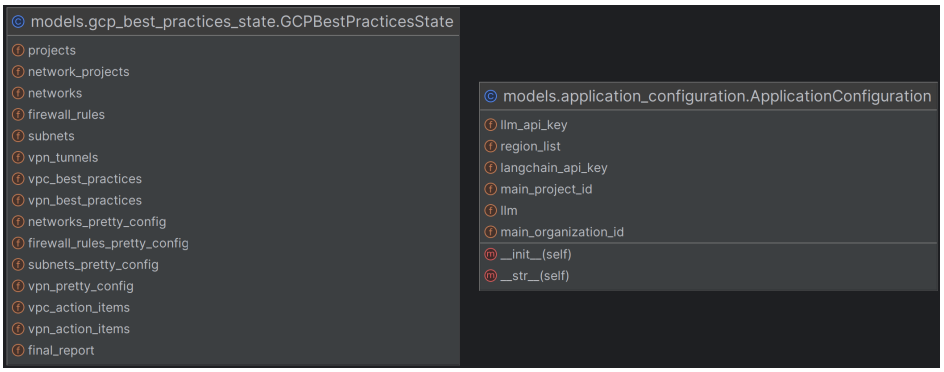


Figura **2-3**.: Modelos utilizados dentro de la aplicación

2.2.2.3. Utilitarios de Aplicación

El módulo de utilitarios de aplicación (ver figura **2-4**) reúne clases que ofrecen métodos de soporte funcional en diversas etapas del flujo principal de la aplicación. A continuación, se describen las clases incluidas:

- `AppUtils`: Esta clase contiene métodos de utilidad para diversas fases de ejecución. Permite, por ejemplo, extraer texto HTML de una página web a partir de su URL y cuenta con un método para obtener

un `api_key`, ya sea desde la consola o del identificador del proyecto principal en Google Cloud almacenado en la clase `ApplicationConfiguration`. Además, configura el seguimiento de la ejecución del agente mediante LangSmith, cuyo uso es opcional.

- `ConsoleUtils`: Dado que la interacción con el aplicativo se realiza principalmente a través de la consola, esta clase centraliza métodos de utilidad para imprimir y leer mensajes, ofreciendo una interfaz estándar y organizada para estos mensajes.

- `InitUtils`: Proporciona métodos específicos para la inicialización de la aplicación. Aquí se configura el sistema de *logging* para gestionar la salida de mensajes en consola, se muestra el mensaje de bienvenida, y, fundamentalmente, se inicializa el objeto `ApplicationConfiguration`. Este objeto almacena los resultados de las tareas clave de la estrategia, tales como Crear conexión a la Organización de Google Cloud 2.1.3.1 e Inicializar el proceso de obtención de la `API key` del modelo LLM 2.1.4.1.

- `ReportUtils`: Esta clase contiene métodos para procesar los resultados del análisis y generar el reporte especificado en la tarea 2.1.9. Dado que el reporte debe ser claro y amigable para el usuario final, `ReportUtils` genera archivos en formatos *markdown* y PDF para facilitar la lectura y presentación de los resultados del análisis.

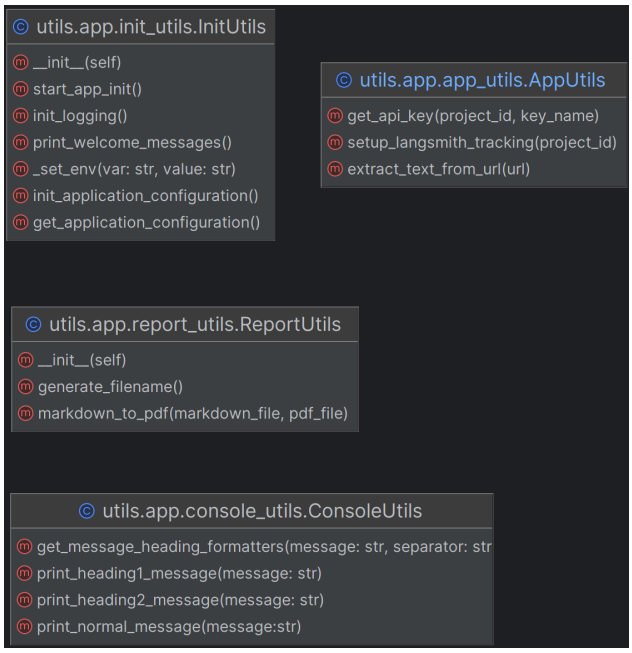


Figura **2-4**.: Clases utilitarias de aplicación

2.2.2.4. Utilitarios de Google Cloud

Los utilitarios de Google Cloud agrupan clases y métodos que ejecutan acciones sobre Google Cloud (ver figura **2-5**) con el fin de obtener la información requerida para ser analizada. Los utilitarios de Google Cloud contienen las clases a continuación:

- `NetworkUtils`: La clase utilitaria de red se encarga de obtener información de VPC, reglas de firewall, subredes de Google Cloud e información de VPN. Esta clase tiene como objetivo ser invocada por demás partes de la aplicación con el fin de obtener la información antes mencionada. Esta clase se encarga de implementar la tarea mencionada en la estrategia 2.1.3.3.
- `OrganizationUtils`: Esta clase se encarga de obtener el id de organización (tarea 2.1.3.2), obtener el proyecto principal que se encuentra dentro de la organización (el proyecto principal es usado para escanear los secretos de como el API Key del LLM y demás así almacenar los proyetos de forma segura), obtener las regiones dispoinibles que existen en Google Cloud, obtener las carpetas de la organización y listar los demás proyectos de la organización.
- `SecretManagerUtils`: Esta clase utilitaria se encarga de obtener los secretos del servicio de Secret Manager, si el usuario no quiere proporcionar el API Key en la consola de ejecución entonces tiene la opción de leerlo desde el servicio de secret manager de Google Cloud desde el identificador principal del proyecto, que se obtuvo mediante la clase utilitaria `OrganizationUtils`.

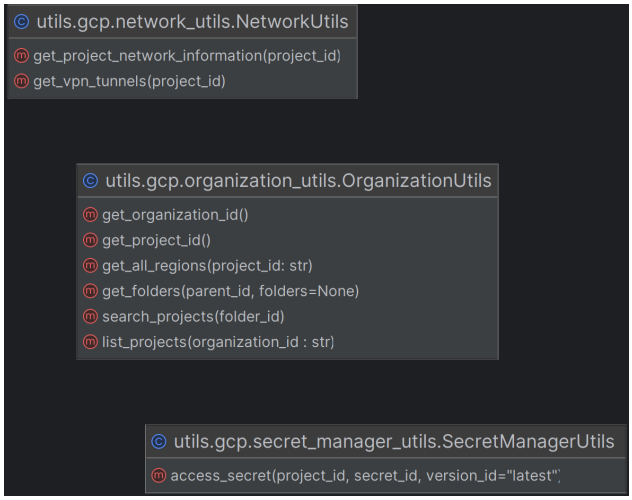


Figura **2-5**.: Clases utilitarias para Google Cloud

2.2.2.5. Agente analizador de mejores prácticas

Este módulo contiene la clase `BestPracticesAnalyzer`, que desempeña una función clave en la ejecución del análisis de buenas prácticas en Google Cloud. Aunque otros módulos no se exploran en profundidad,

en este caso se proporciona un desglose detallado de los métodos de `BestPracticesAnalyzer`, ya que esta clase es el núcleo del análisis del aplicativo. En la figura **2-6**, se incluyen dos gráficos que ilustran su funcionamiento: la figura **2-6**a presenta el diagrama de clases con los métodos y atributos de la clase, mientras que la figura **2-6**b muestra el diagrama de flujo que sigue el agente, creado mediante LangGraph (sección IV.5), y destaca los métodos de la clase que son invocados en este flujo. A continuación, se detallan los métodos de `BestPracticesAnalyzer`:

- `llm`: Atributo que inicializa y almacena el modelo de lenguaje invocado por el agente. En este trabajo se utiliza el modelo *OpenAI 4o*, cumpliendo las tareas de inicialización de API key y del modelo LLM descritas en las secciones 2.1.4.1, 2.1.4.2 y 2.1.4.3.

- `start_analysis`: Método invocado por la clase `App` para iniciar la ejecución del aplicativo. Implementa la tarea 2.1.2.

- `build_graph`: Agrupa la inicialización del LLM, en línea con la tarea 2.1.4.

- `get_all_projects`: Invoca la clase `OrganizationUtils` para obtener recursivamente los proyectos de la organización. Implementa la tarea 2.1.3.1.

- `get_all_networks`: Utiliza `NetworkUtils` para obtener redes, reglas de firewall y subredes de la organización, y normaliza la información obtenida. Implementa las tareas 2.1.3.3 y 2.1.3.4.

- `get_all_vpn_tunnels`: Invoca `NetworkUtils` para recopilar túneles VPN de los proyectos con redes de VPC, normalizando la información. Implementa las tareas 2.1.3.3 y 2.1.3.4.

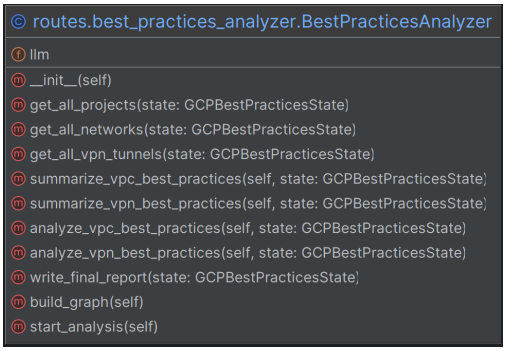
- `summarize_vpc_best_practices`: Obtiene y resume las buenas prácticas de VPC, basándose en la documentación pública de Google Cloud (GCP VPC Best Practices, 2024). Cumple la tarea 2.1.4.4.

- `summarize_vpn_best_practices`: Similar al método anterior, resume las buenas prácticas de VPN usando el LLM. Implementa la tarea 2.1.4.4.

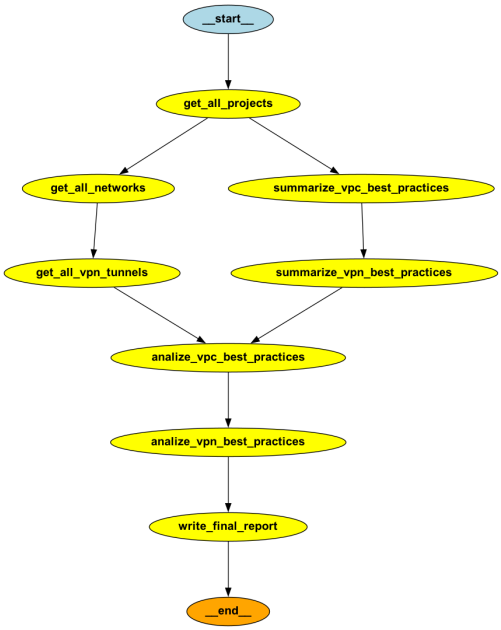
- `analize_vpc_best_practices`: Invoca el LLM para analizar las buenas prácticas de VPC con la información de Google Cloud y la documentación resumida. Implementa las tareas 2.1.5 y 2.1.6.

- `analize_vpn_best_practices`: Similar al anterior, analiza las buenas prácticas de VPN utilizando la información del modelo LLM. Implementa las tareas 2.1.7 y 2.1.8.

- `write_final_report`: Genera un reporte en markdown que sintetiza las recomendaciones del LLM y facilita su revisión por el usuario. Implementa las tareas 2.1.9 y 2.1.10.



(a) Clase `BestPracticesAnalyzer`



(b) Diagrama de flujo de agente creado por la clase `BestPracticesAnalyzer`

Figura **2-6**.: Elementos de la clase principal de flujo de ejecución `BestPracticesAnalyzer`

2.2.3. Flujo de ejecución de la herramienta

2.2.3.1. Prerrequisitos técnicos

La aplicación fue desarrollada principalmente en Python, por lo cual debe ejecutarse en un ordenador compatible con este lenguaje. Adicionalmente, para conectarse a Google Cloud, la aplicación utiliza el SDK de Google Cloud, por lo que se requieren los siguientes prerrequisitos técnicos:

- **Sistema operativo**: Windows, macOS o Linux, en cualquier ordenador capaz de ejecutar Python y con conexión a internet.
- **Python**: Es necesario que Python esté instalado en el ordenador. La versión recomendada para este aplicativo es Python 3.10.x o superior. Las versiones anteriores podrían no garantizar el funcionamiento. La instalación de Python se puede realizar desde su sitio web oficial: `https://www.python.org/downloads/`.
- **pip**: Si el administrador de paquetes de Python `pip` no se ha instalado junto con Python, el usuario debe asegurarse de contar con esta herramienta para gestionar las dependencias del proyecto. En caso de no estar instalado, se puede instalar siguiendo las instrucciones disponibles en `https://pip.pypa.io/en/stable/installation/`.

- **Google Cloud SDK**: El SDK de Google debe estar instalado para establecer conexión con Google Cloud. Las instrucciones de instalación están disponibles en: `https://cloud.google.com/sdk/docs/install`.
- **API Key del modelo de OpenAI**: Para invocar el modelo de lenguaje de OpenAI, es necesario generar y proporcionar una API Key. Esta se puede obtener en `https://platform.openai.com/api-keys`. No se recomienda incluir la API Key en el código fuente; en su lugar, debe ser suministrada en tiempo de ejecución para evitar riesgos de seguridad y costos inesperados.
- **Graphviz**: Esta herramienta es utilizada por Python para generar el grafo de ejecución, como se muestra en la figura **2-6**b.
- **(Opcional) API Key de LangSmith**: Si el usuario desea rastrear las decisiones del agente en la nube mediante LangSmith (ver sección IV.6), puede proporcionar una API Key específica para hacer el seguimiento de la ejecución. La API Key se puede crear en `https://docs.smith.langchain.com/administration/how_to_guides/organization_management/create_account_api_key`. Al igual que con la API Key de OpenAI, se recomienda no incluirla en el código fuente y proporcionarla en tiempo de ejecución.

Es imprescindible contar con todas estas herramientas instaladas para evitar fallos en la ejecución del aplicativo.

2.2.3.2. Flujo de Ejecución

En esta sección se describe el flujo de ejecución de la herramienta el cual presenta cómo se debe correr la herramienta y los resultados obtenidos.

Inicialización

Las instrucciones presentes en esta sección deben ejecutarse una única vez y son requeridas como prerrequisito para la ejecución. Una vez el usuario las ejecute no es necesario ejecutarlas de nuevo, a menos que alguno de los valores proporcionados deba cambiar, en dado caso las instrucciones deben ejecutarse de nuevo:

1. El primer paso es el de conectarse a la consola de Google Cloud para obtener la información que se va a analizar por lo tanto una vez instalado el Google SDK es necesario realizar la inicialización y autorización de la herramienta mediante la guía `https://cloud.google.com/sdk/docs/authorizing#auth-login`.

   a) Inicializar el SDK de google cloud:

```
1  gcloud init
```

Bloque de código 2.1: Inicialización del SDK de Google Cloud.

b) Hacer login mediante una cuenta que tenga como mínimo los siguientes roles de Google Clloud Platform `https://cloud.google.com/iam/docs/understanding-roles`:

c) **Organization administrator:** Acceso para administrar políticas de IAM y ver políticas de la organización para organizaciones, carpetas y proyectos.

d) **Folder Admin:** Proporciona todos los permisos disponibles para trabajar con carpetas.

e) **Network managerment Admin:** Acceso completo a los recursos de gestión de red.

f) **Secret Manager Admin**: Acceso completo para administrar los recursos de Secret Manager.

La ausencia de alguno de estos roles pueden causar problemas en la ejecución del aplicativo, por lo tanto se recomienda tener todos los roles mencionados anteriormente, una vez se tenga un suuario con este rol se debe lanzar el proceso de login mediante el comando:

```
1                gcloud auth login
```

Bloque de código 2.2: Inicialización de usuario del SDK de Google Cloud.

2. Posteriormente es necesario autenticarse con el mismo usuario mencionado anteriormente pero para que la aplicación lea esas credenciales, es decir este login es necesario para que cuando la aplicación de Python invoque el SDK pueda usar las credenciales y los permisos de esa cuenta para llamar el API:

```
1            gcloud auth application-default login
```

Bloque de código 2.3: Inicialización de credenciales de usuario para aplicación en el SDK de Google Cloud.

3. Adicionalmente, es requerida la insalación de las dependencias de python que el programa requiere para su ejecución las cuales se instalan mediante el manejador de paquetes de python. Para lograr este objetivo es necesario clonar el repositorio de GitHub y/o descargar el código fuente `https://github.com/roger8849/gcp_infra_best_practices_analyzer/tree/main/gcp_configurations_analyzer`, Una vez descargado se debe ir a la carpeta `gcp_configurations_analyzer` donde se debe ejecutar el comando 2.4:

4. Adicionalmente, es necesario instalar las dependencias de Python requeridas para la ejecución del programa. Estas dependencias se gestionan mediante el administrador de paquetes de Python `pip`. Para instalar las dependencias, el usuario debe clonar el repositorio de GitHub o descargar el código fuente desde `https://github.com/roger8849/gcp_infra_best_practices_analyzer/tree/main/gcp_configurations_analyzer`. Una vez descargado el proyecto, se debe navegar a la carpeta `gcp_configurations_analyzer` y ejecutar el siguiente comando para instalar los paquetes especificados en `lst:instalacion_dependencias_pip`:

```
pip -r requirements.txt
```

Bloque de código 2.4: Comando de instalación de paquetes requeridos de Python.

Como se menciona anteriormente este proceso sólo se debe ejecutar una sóla vez, no obstante, en caso que se desee usar otro usuario o el token de autenticación del SDK de Google haya expirado entonces se recomienda ejecutar el proceso de nuevo.

Manejo de secretos

El API key del modelo de OpenAI y el API key opcional de LangSmith, utilizado para el seguimiento de la ejecución del modelo de LangChain, son datos sensibles que deben ser almacenados y gestionados de manera segura. Para ello, se recomienda emplear el servicio Google Cloud Secret Manager (Google Secret Manager, 2024). Estos secretos deben almacenarse en el proyecto principal especificado en la ejecución (2.2.3.2) bajo las siguientes claves: `llm_api_key` para el modelo de OpenAI y `langchain_api_key` para LangSmith.

Como se indicó, el uso de Google Cloud Secret Manager es opcional y solo representa una opción para asegurar estos datos. En caso de no utilizar este servicio, es posible proporcionar los API keys de forma manual durante la ejecución del programa, siempre teniendo en cuenta la importancia de proteger la información sensible.

Ejecución

Inicialmente es necesario ejecutar el archivo `app.py` mediante el uso de python como se ven en el bloque de código 2.5

```
python app.py
```

Bloque de código 2.5: Comando para el lanzamiento del aplicativo.

Al iniciar el aplicativo, se mostrará un mensaje de bienvenida que resalta la necesidad de tener el Google Cloud SDK correctamente inicializado. A continuación, el aplicativo solicitará el primer parámetro: el identificador de la organización, que debe ser un número de 12 dígitos. Este paso se ilustra en el bloque de código 2.6.

```
INFO:root:
INFO:root:
INFO:root:#########################################################################
INFO:root:#                                                                       #
INFO:root:# Welcome to the google cloud best practices analyzer powered by LLMs #
INFO:root:#                                                                       #
INFO:root:#########################################################################
INFO:root:
INFO:root:
INFO:root
    :-----------------------------------------------------------------------------------------------
```

```
INFO:root:- This application works with Google Cloud SDK make sure you have it installed and you have run: -
INFO:root:------------------------------------------------------------------------------------------------------

INFO:root:---------------------------------------
INFO:root:- Initialized gcloud sdk: gcloud init -
INFO:root:---------------------------------------
INFO:root:------------------------------------------------------------------------------------------
INFO:root:- Initialized application default credentials: gcloud auth application-default login -
INFO:root:------------------------------------------------------------------------------------------
INFO:root:--------------------------------------------------
INFO:root:- Otherwise the application won't work properly. -
INFO:root:--------------------------------------------------
INFO:root:
INFO:root:Insert your organization id
INFO:root:
INFO:root:
INFO:root:Insert your organization id, must be a number of 12 digits. I.E: <885157069XXX>
INFO:root:
```

Bloque de código 2.6: Mensaje de Bienvenida y requisición del identificador de organización.

Tras ingresar el identificador de la organización, el aplicativo solicitará el identificador principal del proyecto (desde el cual se obtendrán los secretos del Secret Manager), el API key de OpenAI y, a continuación, consultará al usuario si desea realizar el seguimiento de la ejecución del agente mediante LangSmith. Estos parámetros están ilustrados en el bloque de código 2.7.

```

# Obteción del identificador principal de proyecto para obtención de secretos.
Enter a project ID (4-30 chars, lowercase, digits, hyphens): >? project-id

# Se pregunta si el API key de Open AI se obtiene desde el manejador de secretos de Google.
Enter your choice (1 or 2): INFO:root:Choose how to retrieve the llm_api_key:
INFO:root:1. From Google Cloud Secret Manager (project_id: project-id, secret_id: llm_api_key)
INFO:root:2. From the console
1
DEBUG:google.auth._default:Checking None for explicit credentials as part of auth process...
DEBUG:google.auth._default:Checking Cloud SDK credentials as part of auth process...
DEBUG:google.auth.transport.requests:Making request: POST https://oauth2.googleapis.com/token
DEBUG:urllib3.connectionpool:Starting new HTTPS connection (1): oauth2.googleapis.com:443
DEBUG:urllib3.connectionpool:https://oauth2.googleapis.com:443 "POST /token HTTP/11" 200 None

# Se pregunta si se desea hacer seguimiento del aplicativo mediante el uso de Langsmith
Do you want to track application execution with LangSmith? (y/n): >? y
# Si se requiere hacer seguimiento del aplicativo mediante LangSmith se pregunta por el API Key.
```

```
45 Enter your choice (1 or 2): INFO:root:Choose how to retrieve the langchain_api_key:
46 INFO:root:1. From Google Cloud Secret Manager (project_id: project-id, secret_id:
       langchain_api_key)
47 INFO:root:2. From the console
48 1
```

Bloque de código 2.7: Inserción del identiifcador principal del proyecto api key de Open AI y el api key de langmisth respectivamente.

Después de ingresar todos los parámetros requeridos, el aplicativo inicia el proceso de ejecución y realiza un escaneo de las redes VPC, VPN, y una evaluación de sus buenas prácticas. **Es importante señalar que el tiempo de ejecución puede exceder los 5 minutos**, dependiendo del tamaño de la organización y del número de proyectos que se analizarán. Al finalizar el análisis, el aplicativo mostrará un mensaje de conclusión, indicando la ubicación del reporte final generado, tal como se observa en el bloque de código 2.8.

```
100 INFO:root:
101 INFO:root:################################################################################
102 INFO:root:#                                                                              #
103 INFO:root:# Analysis ended find the full report at                                      #
104 INFO:root:# /Users/rramirezespejo/dev_home/gcp_infra_best_practices_analyzer/           #
105 INFO:root:# gcp_configurations_analyzer/reports/final_report_20241114-095019.md #
106 INFO:root:#                                                                              #
107 INFO:root:################################################################################
108 INFO:root:
```

Bloque de código 2.8: Mensaje final del aplicativo post-ejecución.

Finalmente, si el rastreo mediante LangSmith fue configurado y ejecutado, los resultados de la ejecución del agente pueden visualizarse en `https://smith.langchain.com/`. En el proyecto configurado, es posible observar detalles de la ejecución, como se muestra en la figura **2-7**. Por ejemplo, se puede ver que todos los métodos del grafo representado en el diagrama de flujo de la figura **2-6**b fueron ejecutados, y que el tiempo total de ejecución del agente fue de 354.43 segundos.

Visualización de resultados

Los resultados de cada ejecución se guardan en un archivo de formato *Markdown* (Markdown Guide, 2024), un estándar ampliamente utilizado para visualización. Al finalizar cada ejecución exitosa, el programa genera un reporte que se almacena en la carpeta `reports` con el siguiente formato de nombre de archivo: `final_report_<timestr>.md`, donde `timestr` sigue la convención `%Y%m%d-%H%M%S`. Por ejemplo, un archivo generado podría tener el nombre `final_report_20241101-154503.md`. Puede consultarse un ejemplo de estos reportes en el repositorio de GitHub en la ruta `https://github.com/roger8849/gcp_infra_best_practices_analyzer/tree/main/gcp_configurations_analyzer/reports`.

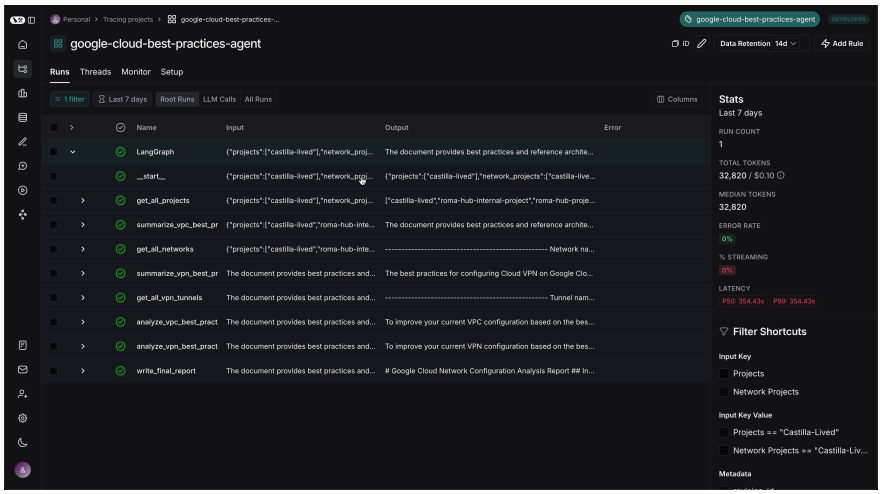


Figura **2-7**.: Visualización de resultados de rastreo de ejecución de agente en LangSmith

3. Evaluación de la herramienta de análisis de buenas prácticas de Google Cloud

Este capítulo expone los resultados de la herramienta desarrollada y descrita a lo largo del capítulo 2. En la sección 3.1 se presenta la estructura de la organización de Google Cloud utilizada para la evaluación, y posteriormente, en la sección 3.2, se detallan los resultados obtenidos tras ejecutar la herramienta sobre esta organización.

3.1. Estudio de caso: Organización de prueba de Google Cloud

En esta sección se presenta el estudio de caso que permite evaluar la aplicación desarrollada. Para este fin, se creó una organización de Google Cloud bajo el dominio *roma.joonix.net*, con la estructura de carpetas y proyectos que se muestra en la figura **3-1**. La evaluación no se realizó sobre una organización real para evitar la exposición de datos sensibles de configuración, y debido a la imposibilidad de acordar con un cliente real el uso de datos para la experimentación.
No obstante, la configuración realizada para esta organización de prueba cumple con estándares de alta calidad y sigue el principio arquitectónico y la arquitectura de referencia de red y organización de *hub y spokes* como lo dicta la documentación oficial de Google Cloud (Google Cloud Architecture Center, 2024). Para evaluar el análisis en esta organización, se crearon diversos recursos de red, incluyendo VPCs, subredes, reglas de firewall y redes VPN. Este diseño, que se puede observar en la figura **3-2**, permitió al aplicativo realizar la evaluación de forma completa y eficiente.
Tomando como referencia los diagramas de las figuras **3-1** y **3-2**, la organización utilizada para la prueba incluye los proyectos clave que se detallan a continuación. Cabe destacar que esta lista no es exhaustiva, sino que se centra en aquellos proyectos relevantes para la topología de red y sobre los cuales se enfoca el análisis del aplicativo.

- **Organización roma.joonix.net**: Este dominio no pertenece a ninguna organización real; sin embargo, es un dominio registrado en el DNS global y fue validado dentro de Google Cloud.
- **Proyecto de red roma-hub-project**: Este proyecto contiene una VPC con cuatro subredes en diferentes regiones y tres reglas de firewall. Además, al ser el proyecto HUB (ver anexo C.1) de conectividad,

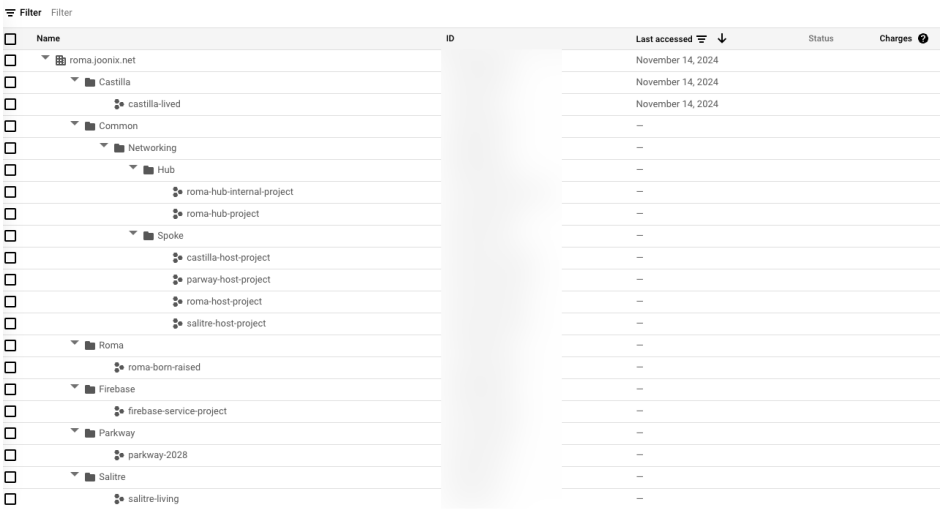


Figura **3-1**.: Organización de proyectos dentro de la organización roma.joonix.net. Los identificadores de estos proyectos fueron difuminados intencionalmente para no exponer los identificadores reales de la organización.

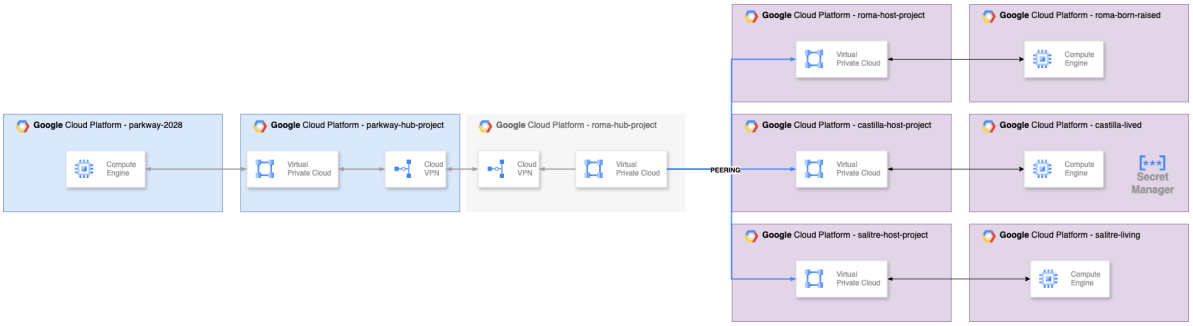


Figura **3-2**.: Topología de red de organización de prueba.

cuenta con 11 túneles de VPN que simulan conexiones hacia otros proyectos en Google Cloud y otras nubes. Este proyecto es crucial, ya que centraliza gran parte de la topología de red.

- **Proyectos Host-Spoke**: Los proyectos spoke (ver anexo C.2) están conectados al proyecto de conectividad mediante peering, lo que permite que las distintas redes VPC puedan interconectarse. Estos proyectos están destinados a almacenar configuraciones de red para las aplicaciones y servicios. Así, las configuraciones de red se encuentran en los proyectos Spoke, mientras que las cargas de trabajo están en los proyectos de servicio, promoviendo una división de responsabilidades y asegurando que los desarrolladores o propietarios de los proyectos de servicio no realicen configuraciones de red que puedan afectar otras aplicaciones. Los proyectos host-spoke de la organización de prueba son:
  - *castilla-host-project, roma-host-project y salitre-host-project:* Estos proyectos representan unidades de negocio dentro de la organización y almacenan sus configuraciones de red. Están conec-

tados mediante peering al proyecto hub, lo cual asegura que el tráfico entre Google Cloud y otras nubes sea canalizado a través del *roma-hub-project* de manera centralizada y en conformidad con las buenas prácticas.

- *parkway-host-project:* Aunque este es un proyecto de Google Cloud, fue creado para simular una conexión hacia un entorno on-premises o en otra nube. Por lo tanto, la conexión se realiza mediante una VPN de alta disponibilidad en lugar de un peering entre proyectos, replicando el escenario previamente descrito.

- **Proyectos de servicio**: Estos proyectos de servicio utilizan las redes configuradas en los proyectos Spoke y no deben contener configuraciones de red propias, ya que estas deben ser administradas por el equipo de red de la compañía. Los proyectos *castilla-lived, roma-born-raised, parkway-2028 y salitre-living* emplean las configuraciones de red de los proyectos host mencionados anteriormente. En particular, el proyecto *castilla-lived* contiene los secretos configurados para el API key de Open AI y el API key de langsmith, lo cual ejemplifica que estos proyectos deben enfocarse en las cargas de trabajo y no en configuraciones de red.

En la sección 3.2 se presentan en detalle los elementos analizados por el aplicativo. Este análisis tiene como objetivo identificar y evaluar los servicios y configuraciones implementados en la organización de prueba, permitiendo así determinar el estado de las configuraciones realizadas y su alineación con las buenas prácticas.

3.2. Análisis y resultados

Basándose en la organización de prueba y el software descrito, se presentan los resultados obtenidos durante la ejecución de las pruebas. Estos resultados se encuentran documentados en un archivo de resultados almacenado en el repositorio de GitHub. Dicho archivo puede visualizarse directamente en la plataforma GitHub o mediante cualquier editor compatible con Markdown. Aunque el uso de un visualizador de Markdown es opcional, este facilita una experiencia más agradable y estructurada para el usuario final.

En las subsecciones posteriores de este documento se analizan y describen las principales características y hallazgos derivados de las pruebas realizadas. Para consultar el informe completo y todos los detalles de los resultados, el lector puede acceder al archivo público disponible en GitHub mediante el siguiente enlace: `https://github.com/roger8849/gcp_infra_best_practices_analyzer/blob/main/gcp_configurations_analyzer/reports/final_report_20241115-221410.md`.

3.2.1. Sección introductoria del reporte

El reporte incluye una sección introductoria, como se muestra en la figura **3-3**. En esta sección, se explica que el reporte está dividido en tres partes principales: Configuraciones de red (3.2.2), Resúmen de buenas prácticas (3.2.3), que presenta un resumen de las prácticas recomendadas implementadas o pendientes y Análisis de las recomendaciones (3.2.4).

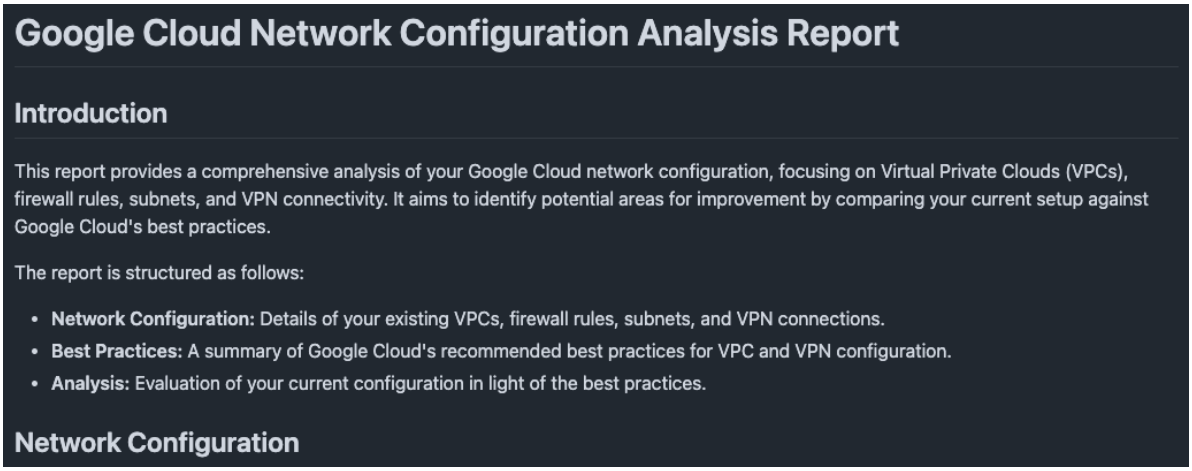


Figura **3-3**.: Sección introductoria de reporte de resultados.

3.2.2. Información de red obtenida de Google Cloud

Dado que el análisis se centra en la configuración de Google Cloud, se considera esencial que el reporte incluya detalles sobre las reglas de VPC (3.1), las reglas de firewall (3.2), las subredes (3.3) y las conexiones VPN (3.4) detectadas en la infraestructura de Google Cloud.

```
1  ------------------------------------------------
2  Network name: vpc-roma-internal-hub
3  Auto create subnetworks: False
4  Network creation timestamp: 2024-09-12T09:33:50.505-07:00
5  Network description:
6  Network Enable ULA Internal IPv6: False
7  Network Firewall Policy:
8  Network Gateway:
9  Network Ipv4 Range:
10 Network id: 3529097084580833537
11 Network Internal IPv6 range:
12 Network kind: compute#network
13 Network MTU: 0
14 Network Firewall Policy enforcer: AFTER_CLASSIC_FIREWALL
15 Network Peerings: []
```

```
Network Routing config: routing_mode: "GLOBAL"

------------------------------------------------
------------------------------------------------
Network name: vpc-roma-hub
Auto create subnetworks: False
Network creation timestamp: 2024-08-15T13:38:23.969-07:00
Network description:
Network Enable ULA Internal IPv6: False
Network Firewall Policy:
Network Gateway:
Network Ipv4 Range:
Network id: 8688276164912077264
Network Internal IPv6 range:
Network kind: compute#network
Network MTU: 0
Network Firewall Policy enforcer: AFTER_CLASSIC_FIREWALL
Network Peerings: []
Network Routing config: routing_mode: "GLOBAL"

------------------------------------------------
```

Bloque de código 3.1: Ejemplo de configuraciones de red obtenidas desde Google Cloud.

```
------------------------------------------------
Firewall rule id: 715227665512627571
Firewall rule name: allow-all
Firewall rule description:
Firewall rule priority: 1000
Firewall rule source ranges: ['0.0.0.0/0']
Firewall rule source service accounts: []
Firewall rule target tags: []
Firewall rule target service accounts: []
Firewall rule kindcompute#firewall
Firewall rule disabled: False
------------------------------------------------
------------------------------------------------
Firewall rule id: 8205348981296712063
Firewall rule name: deny-all2
Firewall rule description:
Firewall rule priority: 65535
Firewall rule source ranges: ['0.0.0.0/0']
Firewall rule source service accounts: []
Firewall rule target tags: []
Firewall rule target service accounts: []
```

```
Firewall rule kindcompute#firewall
Firewall rule disabled: False
------------------------------------------------
```

Bloque de código 3.2: Ejemplo de configuraciones de reglas de firewall obtenidas desde Google Cloud.

```
------------------------------------------------
Subnet ip_ciddr range10.0.0.0/24
Subnet purposePRIVATE
Subnet secondary ip range[]
Subnet internal ipv6 prefix
Subnet external ipv6 prefix
------------------------------------------------
------------------------------------------------
Subnet ip_ciddr range172.16.20.16/28
Subnet purposePRIVATE
Subnet secondary ip range[]
Subnet internal ipv6 prefix
Subnet external ipv6 prefix
------------------------------------------------
```

Bloque de código 3.3: Ejemplo de configuraciones de subredes obtenidas desde Google Cloud.

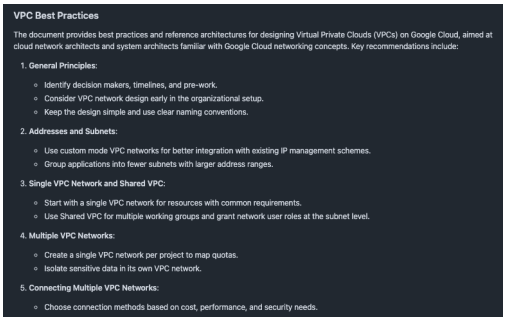
```
------------------------------------------------
Tunnel name ha-vpn-spoke-hub-southamerica-west1-cl-noprod-ha-tunnel1
Tunnel IKE Version: 2
Tunnel Peer ip: 34.153.33.202
Tunnel status: ESTABLISHED
Tunnel Shared secret: *************
Tunnel Kind: compute#vpnTunnel
------------------------------------------------
------------------------------------------------
Tunnel name vpn-to-oracle-tunnel1
Tunnel IKE Version: 2
Tunnel Peer ip: 141.148.64.14
Tunnel status: ESTABLISHED
Tunnel Shared secret: *************
Tunnel Peer GCP Gateway:
Tunnel Kind: compute#vpnTunnel
------------------------------------------------
```

Bloque de código 3.4: Ejemplo de configuraciones de VPN obtenidas desde Google Cloud.

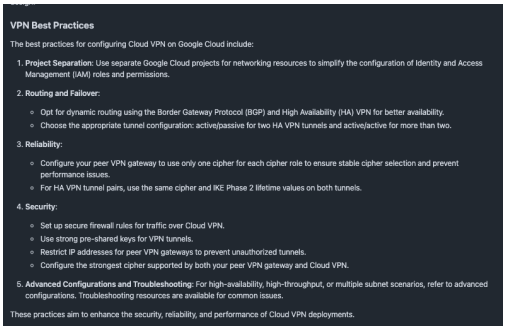
A partir de esta información, el aplicativo genera un reporte que presenta las buenas prácticas identificadas y resumidas a partir de la documentación pública de Google Cloud, como se describe en la sección 3.2.3.

3.2.3. Extracción de buenas prácticas de documentación de Google Cloud

Tal como se observa en las figuras **3-4**a y **3-4**b, el reporte generado por el aplicativo incluye las buenas prácticas obtenidas desde la documentación pública de Google Cloud.



(a) Resumen de buenas prácticas de VPC incluidas en el reporte final



(b) Resumen de buenas prácticas de VPN incluidas en el reporte final

Figura **3-4**.: Resumen de documentación de buenas prácticas obtenida de la documentación oficia de Google Cloud.

Como se puede ver, el agente sintetiza el contenido considerablemente, destacando los puntos más relevantes de la documentación para proporcionar un resumen más accesible y fácil de digerir para el usuario final. Este paso demuestra la utilidad del aplicativo para instruir al usuario en las mejores prácticas que se analizarán en detalle más adelante en la sección 3.2.4.

3.2.4. Reporte de Análisis

Una vez que el aplicativo y el agente recopilan la información requerida de Google Cloud junto con las buenas prácticas correspondientes, el aplicativo solicita al agente, y por ende al modelo, que analice las mejores prácticas aplicables a la configuración presentada. Este análisis se detalla en las secciones 3.2.4.1 y 3.2.4.2, respectivamente.

3.2.4.1. Reporte de Análisis de VPC

Teniendo en cuenta la documentación de buenas prácticas extraída, el modelo generó una a una las secciones incluidas en el documento de prácticas de VPC: Principios Generales, Direcciones y Subredes, VPC Única y VPC Compartida, Múltiples Redes de VPC, Conexión de Múltiples Redes de VPC, Seguridad en la Red, Servicios de Red, Acceso a VPC mediante API, Registro, Monitoreo y Visibilidad, y Consideraciones Adicionales. En el bloque de código 3.5, las recomendaciones tuvieron en cuenta la configuración recibida de VPC. Por ejemplo, en la sección de Principios Generales, el modelo sugiere que las redes `vpc-roma-internal-hub` y `vpc-roma-hub` podrían unirse para simplificar el diseño de red, lo cual tiene sentido, ya que las

dos redes tienen nombres muy similares y unirlas reduciría la administración. Por otro lado, en la sección de Seguridad, el modelo destaca que existe una regla de firewall que permite todo el tráfico (`allow-all`) hacia la red, lo que representa una brecha de seguridad.

Adicionalmente, el modelo sugiere que las redes `vpc-roma-hub` y `vpc-castilla-host` podrían requerir conexión mediante VPN o Peering. Sin embargo, esta conexión ya existe, pero ese contexto no se le proporcionó al modelo, por lo que faltó alimentar ese contexto para obtener una recomendación más precisa.

```
### General Principles
1. **Simplify Network Design**:
   - Consider consolidating VPC networks where possible to reduce complexity. For example,
       evaluate if `vpc-roma-internal-hub` and `vpc-roma-hub` can be merged if they serve similar
        purposes.

### Connecting Multiple VPC Networks
7. **VPC Peering and Connectivity**:
   - Consider using VPC Network Peering or Cloud VPN for connecting networks that need to
       communicate. For example, if `vpc-roma-hub` and `vpc-castilla-host` need to communicate,
       evaluate the best method based on cost, performance, and security.

### Network Security
8. **Firewall Rules**:
   - Review and refine firewall rules. For instance, the `allow-all` rules in `vpc-castilla-host`
        and `vpc-salitre-host` are broad and could pose security risks. Replace them with more
       specific rules that only allow necessary traffic.
```

Bloque de código 3.5: Recomendaciones específicas con base en la configuración definida en Google Cloud acerca con respecto a las redes de VPC.

En segunda instancia, algunas de las recomendaciones presentes en el bloque de código 3.6 son más genéricas, pero intentan dar pistas sobre la configuración requerida. Por ejemplo, la primera recomendación sobre convenciones de nombres sugiere que las redes, subredes y reglas de firewall deberían tener una descripción que proporcione metadatos sobre su propósito, facilitando así su administración. Además, el modelo resalta que todas las VPC analizadas usan el *modo personalizado*, lo cual es considerado una buena práctica, indicando al usuario que debe mantener esta configuración.

Por último, la recomendación sobre el uso de VPCs compartidas debe ser considerada; sin embargo, todas las redes VPC host son redes compartidas y ya cumplen con esta configuración. No obstante, este contexto no se le proporcionó al modelo, lo que le impidió inferir correctamente que esta recomendación ya estaba satisfecha.

```
### General Principles
2. **Naming Conventions**:
   - Ensure all networks, subnets, and firewall rules have descriptive names and descriptions to
       improve manageability and clarity.
```

```
4
5  4. **Custom Mode VPC**:
6     - You are already using custom mode VPCs, which is good for integration with existing IP
          management schemes. Ensure that the IP ranges do not overlap with other networks,
          especially if planning to connect them.
7
8  ### Single VPC Network and Shared VPC
9  5. **Shared VPC**:
10    - If multiple projects require access to shared resources, consider using a Shared VPC to
          centralize network management and improve security controls.
```

Bloque de código 3.6: Recomendaciones genéricas que no tienen en cuenta configuración definida en Google Cloud acerca con respecto a las redes de VPC.

Una vez completado el análisis de VPC, el aplicativo procede a generar las buenas prácticas de VPN y las agrega al reporte en la sección 3.2.4.2.

3.2.4.2. Reporte de Análisis de VPN

En línea con el análisis realizado en la sección 3.2.4.1, el modelo utilizó una estrategia similar para generar recomendaciones basadas en las secciones más relevantes de la documentación de buenas prácticas de VPN (GCP VPN Best Practices, 2024). El modelo estructuró las recomendaciones en las siguientes categorías: Separación de proyectos, Enrutamiento y conmutación por error, Fiabilidad, Seguridad, Alta disponibilidad y rendimiento, Monitoreo y alertas, y Revisión de documentación.

En el bloque de código 3.7, el modelo incluyó una recomendación específica destacable: detectó que los túneles configurados deberían estar en una configuración activo-activo en regiones separadas. Esta recomendación busca mejorar la conmutación por error y la resiliencia de la infraestructura.

```
1  Routing and Failover:
2
3  2. **Routing and Failover**:
4     - **Tunnel Configuration**: You have multiple HA VPN tunnels. Ensure that you are using active
          /active configurations for tunnels where more than two are present, such as in your `us-
          central1-dev` and `us-east4-cl-prod` regions.
```

Bloque de código 3.7: Recomendaciones específicas con base en la configuración definida en Google Cloud acerca con respecto a las redes de VPN.

En una segunda instancia, el modelo no logró generar más recomendaciones específicas y se centró en sugerencias más generales derivadas de la documentación, como se detalla en el bloque de código 3.8.

La primera recomendación sugiere que las configuraciones de VPN deberían residir en proyectos separados, lo cual ya se cumple, ya que estas configuraciones están contenidas en el proyecto `roma-hub-project`. La segunda recomendación propone la creación de túneles adicionales para mejorar el rendimiento de la VPN,

aunque esta sugerencia es genérica y debería fundamentarse en métricas de rendimiento que evalúen el tráfico actual en los túneles; por lo tanto, su aplicabilidad es subjetiva. Finalmente, la última recomendación aboga por documentar las configuraciones de VPN para alinear las prácticas con los objetivos organizacionales y las mejores prácticas. Esta sugerencia, sin embargo, proviene únicamente de la documentación y no refleja un análisis detallado de la configuración actual.

```
1
2  1. **Project Separation**:
3     - Ensure that your VPN configurations are organized within separate Google Cloud projects for
          different environments (e.g., production, non-production) to simplify IAM role management
          and enhance security.
4
5
6  5. **High Availability and Throughput**:
7     - For regions with high traffic, consider advanced configurations to support high availability
           and throughput. This might include additional tunnels or optimizing existing
          configurations for better performance.
8
9  7. **Documentation and Review**:
10    - Document your current VPN configurations and regularly review them to ensure they align with
           best practices and organizational policies. This will help in maintaining consistency and
           security across your network.
```

Bloque de código 3.8: Recomendaciones genéricas que no tienen en cuenta configuración definida en Google Cloud acerca con respecto a las redes de VPN.

Las recomendaciones genéricas presentadas en el bloque de código 3.8 muestran que el modelo, a pesar de no contar con información completa, intenta generar sugerencias basadas exclusivamente en la documentación. Esto evidencia la necesidad de proporcionar al modelo un contexto más amplio, no limitándose únicamente a las configuraciones de los túneles. Para que el modelo pueda realizar un análisis más acertado, sería esencial incluir información adicional sobre las conexiones, como la configuración de los extremos opuestos de la VPN. Esto permitiría un razonamiento más fundamentado y recomendaciones más específicas y útiles.

Con base en los resultados obtenidos y presentados, se puede concluir que el objetivo del proyecto se ha cumplido de manera satisfactoria. Aunque el aplicativo es susceptible de mejoras, los resultados alcanzados dentro del tiempo y el alcance definidos demuestran que los modelos de lenguaje pueden ser útiles para generar recomendaciones efectivas de configuración. Esto abre la posibilidad de expandir su uso a otros servicios de nube.

Sin embargo, sería interesante explorar un enfoque híbrido que combine el análisis de buenas prácticas basado en modelos de lenguaje con un enfoque más determinista. Esto podría mejorar la precisión y robustez de las recomendaciones, ya que, como se visualizó en el análisis de VPC y VPN, algunas recomendaciones

fueron demasiado genéricas y otras sí tuvieron en cuenta la configuración específica presente en Google Cloud. Para un análisis más detallado de las conclusiones y propuestas de trabajo futuro, se invita al lector a consultar el capítulo de conclusiones 4.

4. Conclusiones y propuesta de trabajo futuro

4.1. Conclusiones y Trabajo Futuro

El presente trabajo ha explorado el uso de modelos de lenguaje de gran escala (LLMs) y agentes de LangChain para realizar un análisis de buenas prácticas en redes de Google Cloud. A través de este enfoque, fue posible implementar una herramienta que facilita la comprensión de las configuraciones actuales y las contrasta con recomendaciones de buenas prácticas. Sin embargo, al basarse exclusivamente en capacidades de generación de texto, como el análisis de contexto, el resumen y procesamiento de documentación y la generación de recomendaciones, los resultados obtenidos presentan ciertas limitaciones. Los análisis generados por el LLM fueron de naturaleza genérica y a menudo carecen de concreción específica, lo cual se debe en parte a la falta de especialización o ajuste del modelo en relación con configuraciones concretas y específicas de Google Cloud.

Este proyecto ha permitido sentar las bases de una herramienta para el análisis de buenas prácticas en Google Cloud, utilizando modelos de lenguaje y agentes para simplificar el acceso y la aplicación de recomendaciones de infraestructura. No obstante, la calidad de los resultados aún puede mejorarse mediante ajustes específicos en los modelos y en los enfoques de análisis. En conjunto, los resultados obtenidos ofrecen una visión de cómo los avances en IA pueden aplicarse para apoyar la gestión y configuración segura de recursos en la nube, abriendo también un espacio para futuras mejoras que profundicen en la precisión, autonomía y viabilidad financiera de estas soluciones.

4.1.1. Limitaciones y Futuro del Trabajo en Fine-Tuning y Modelos Alternativos

Para obtener recomendaciones más precisas y detalladas, se recomienda como trabajo futuro realizar un proceso de ajuste fino (fine-tuning) del modelo, empleando ejemplos de configuraciones específicas que cumplen con las buenas prácticas en lugar de depender exclusivamente de la documentación general. Este enfoque podría permitir al modelo generar recomendaciones más contextualizadas, minimizando la necesidad de que el LLM haga inferencias amplias que podrían afectar la precisión de las recomendaciones. Además, explorar el rendimiento de otros modelos de lenguaje podría enriquecer la calidad de las respuestas y ofrecer diferentes perspectivas en el análisis de configuraciones.

4.1.2. Expansión del Trabajo a Otros Servicios de Google Cloud

Dado que el tiempo de ejecución de este proyecto fue limitado (15 semanas), se priorizó el análisis de servicios fundacionales de red, como VPC y VPN, que, a pesar de su importancia, presentan un vacío en cuanto a la disponibilidad de recomendaciones en Google Recommender. Si bien Google Cloud ofrece una amplia gama de servicios que podrían beneficiarse de una revisión de buenas prácticas, la decisión de centrarse en VPC y VPN se basó en el análisis presentado en la sección 1.2.1, donde se consideraron factores como la falta de cobertura en Google Recommender y el impacto de estos servicios en la arquitectura general de la nube. Como trabajo futuro, sería interesante extender la herramienta para incluir otros servicios clave, como almacenamiento, bases de datos y seguridad, lo que brindaría una visión más completa y asistiría a las organizaciones en la adopción integral de buenas prácticas en Google Cloud. Sin embargo, es crucial considerar que cada servicio posee particularidades que requerirían un tiempo considerable de ajuste y prueba.

4.1.3. Potencial de un Enfoque Mixto: Reglas y Modelos de LLM

Este trabajo se centró en un enfoque puramente basado en LLMs y agentes, lo cual, si bien ofrece una capacidad de análisis general y adaptativa, puede carecer de la precisión de un sistema basado en reglas programáticas específicas para Google Cloud. Como propuesta de mejora, podría explorarse un método mixto que combine reglas programáticas predefinidas y modelos de LLM. Este enfoque mixto podría permitir un análisis más detallado y con menos ambigüedad, utilizando reglas específicas para configuraciones técnicas, mientras que el modelo de lenguaje podría asistir en tareas de análisis, explicación y generación de recomendaciones cuando las reglas no cubran todas las variaciones posibles.

4.1.4. Contribuciones al Uso de IA y Computación en la Nube

Este proyecto destaca cómo herramientas avanzadas de computación en la nube y de inteligencia artificial pueden integrarse para contribuir al proceso de adopción de nubes públicas mediante el uso de agentes y modelos de lenguaje. La adopción de modelos de lenguaje en esta herramienta permite que organizaciones y usuarios realicen un análisis más detallado y accesible de sus configuraciones de Google Cloud, simplificando la interpretación de documentación y recomendaciones. Esta integración representa un aporte importante, ya que se promueve el uso de recursos programáticos para asistir en la implementación de prácticas de seguridad y eficiencia en la nube.

4.1.5. Consideraciones sobre el Costo y la Dependencia de Grandes Modelos de Lenguaje

El uso de modelos de lenguaje presenta un avance significativo en cuanto a la capacidad de generar análisis no deterministas, lo cual es beneficioso en escenarios complejos de toma de decisiones en la nube. No obstante, su uso también plantea un desafío importante en términos de costos y dependencia de terceros. Estos modelos, en su mayoría, son propiedad de grandes empresas tecnológicas, lo cual limita la autonomía de las organizaciones que deseen implementar soluciones avanzadas sin incurrir en costos adicionales. La opción de entrenar modelos propios, aunque factible, requiere una inversión considerable en recursos computacionales y energéticos, lo que vuelve a colocar a las organizaciones en una posición de dependencia frente a proveedores de servicios de nube y de IA. En un contexto de uso más intensivo, este aspecto financiero se convierte en un factor crucial a considerar.

Referencias

Agarwal, A., Siddharth, S., and Bansal, P. (2016). Evolution of cloud computing and related security concerns. *2016 Symposium on Colossal Data Analysis and Networking, CDAN 2016*.

Arif, H., Hajjdiab, H., Harbi, F. A., and Ghazal, M. (2019). A comparison between google cloud service and icloud. *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019*, pages 337–340.

AWS (2024). Cloud computing with aws. `https://aws.amazon.com/what-is-aws/`.

Briganti, G. (2024). How chatgpt works: a mini review. *European Archives of Oto-Rhino-Laryngology*, 281:1565–1569.

Bryman, A. (2016). *Social Research Methods*. Oxford University Press.

Buyya, R., Broberg, J., and Goscinski, A. (2011). Cloud computing: Principles and paradigms. *Cloud Computing: Principles and Paradigms*.

Cao, D. and Jun, W. (2024). Llm-cloudsec: Large language model empowered automatic and deep vulnerability analysis for intelligent clouds. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE.

Cass Information Systems, I. (2023). Why aws trusted advisor fails to live up to its name. `https://www.cassinfo.com/cloud-management-blog/why-aws-trusted-advisor-fails-to-live-up-to-its-name`.

Chang, E. Y. (2023). Examining gpt-4: Capabilities, implications and future directions. In *The 10th International Conference on Computational Science and Computational Intelligence*.

Easin Arafat, M., Asuah, G., Saha, S., and Orosz, T. (2023). Empowering real-time insights through llm, langchain, and sap hana integration. In *The International Conference on Recent Innovations in Computing*, pages 483–495. Springer.

Fagan, F. (2024). A view of how language models will transform law.

Furht, B. and Escalante, A. (2010). *Handbook of Cloud Computing*. Springer US.

Garfinkel, S. L. and Abelson, H. (1999). *Architects of the Information Society: Thirty-Five Years of the Laboratory for Computer Science at MIT*. The MIT Press.

GCP BigQuery (2023). Cloud data warehouse to power your data-driven innovation. `https://cloud.google.com/bigquery/`.

GCP BigQuery Best Practices (2023). Best practices for cloud firestore. `https://cloud.google.com/`

bigquery/docs/best-practices-performance-compute.

GCP Github Repository (2023). Public google cloud platform repository. https://github.com/GoogleCloudPlatform.

GCP GKE Configuration Choices (2023). Abouc cluster configuration choices. https://cloud.google.com/kubernetes-engine/docs/concepts/types-of-clusters.

GCP Looker Studio (2023). Looker studio. https://cloud.google.com/looker-studio?hl=en.

GCP Recommender (2023). Google cloud recommender overview. https://cloud.google.com/recommender/docs/overview.

GCP Service Summary (2023). Google cloud platform services summary. https://cloud.google.com/terms/services.

GCP VPC Best practices (2023). Best practices and reference architectures for vpc design. https://cloud.google.com/architecture/best-practices-vpc-design.

GCP VPC Best Practices (2024). Best practices and reference architectures for vpc design. https://cloud.google.com/architecture/best-practices-vpc-design.

GCP VPN Best Practices (2024). Best practices for cloud vpn. https://cloud.google.com/network-connectivity/docs/vpn/concepts/best-practices.

Google (2024). Explore over 150+ google cloud products. https://cloud.google.com/products?hl=en.

Google Cloud Architecture Center (2024). Hub-and-spoke network architecture. https://cloud.google.com/architecture/deploy-hub-spoke-vpc-network-topology.

Google Cloud Setup (2024). Google cloud setup checklist. https://cloud.google.com/docs/enterprise/setup-checklist.

Google Cloud VPC (2024). Virtual private cloud (vpc) overview. https://cloud.google.com/vpc/docs/overview.

Google Config Validator (2024). Google cloud platform config validator. https://github.com/GoogleCloudPlatform/config-validator.

Google Policy Analyzer (2024). Policy analyzer for allow policies. https://cloud.google.com/policy-intelligence/docs/policy-analyzer-overview.

Google Secret Manager (2024). Secret manager. https://cloud.google.com/secret-manager?hl=es.

Google Service Models (2024). Paas vs. iaas vs. saas vs. caas: How are they different? https://cloud.google.com/learn/paas-vs-iaas-vs-saas.

Google Terraform Validator (2024). Google cloud platform config validator. https://github.com/GoogleCloudPlatform/terraform-validator.

Hart-Davis, G. (2021). *Teach Yourself VISUALLY Google Workspace*. John Wiley & Sons.

Ito, T., Kuribayashi, T., Hidaka, M., Suzuki, J., and Inui, K. (2020). Langsmith: An interactive academic text revision system. *arXiv preprint arXiv:2010.04332*.

Jeong, C. (2024). A study on the implementation method of an agent-based advanced rag system using

graph.

Kwon, S., Lee, S., Kim, T., Ryu, D., and Baik, J. (2023). Exploring llm-based automated repairing of ansible script in edge-cloud infrastructures. *Journal of Web Engineering*, 22(6):889–912.

LangChain (2024). Langchain introduction. `https://python.langchain.com/docs/introduction/`.

LangGraph (2024). Langgraph quick start. `https://langchain-ai.github.io/langgraph/tutorials/introduction/`.

LangSmith (2024). Langsmith introduction. `https://docs.smith.langchain.com/`.

Lin, T., Yan, P., Song, K., Jiang, Z., Kang, Y., Lin, J., Yuan, W., Cao, J., Sun, C., and Liu, X. (2024). Langgfm: A large language model alone can be a powerful graph foundation model. *arXiv preprint arXiv:2410.14961*.

Lütkepohl, H. (2013). *Chapter 6: Vector autoregressive models*. Edward Elgar Publishing, Cheltenham, UK.

Markdown Guide (2024). The markdown guide is a free and open-source reference guide that explains how to use markdown, the simple and easy-to-use markup language you can use to format virtually any document. `https://www.markdownguide.org/`.

Mell, P. M. and Grance, T. (2011). The nist definition of cloud computing.

Menzies, T. and Zimmermann, T. (2013). Software analytics: So what? *IEEE Software*, 30(4):31–37.

Menzies, T. and Zimmermann, T. (2018). Software Analytics: What's Next? *IEEE Software*, 35(5):64–70.

Mohammed Sadeeq, M., Abdulkareem, N. M., Zeebaree, S. R. M., Mikaeel Ahmed, D., Saifullah Sami, A., and Zebari, R. R. (2021). Iot and cloud computing issues, challenges and opportunities: A review. *Qubahan Academic Journal*, 1(2):1–7.

Mukherjee, R., Tripp, O., Liblit, B., and Wilson, M. (2022). Static analysis for aws best practices in python code. *Leibniz International Proceedings in Informatics, LIPIcs*, 222.

Nascimento, E., García, G., Victorio, W., Lemos, M., Izquierdo, Y., Garcia, R., Leme, L., and Casanova, M. A. (2023). A family of natural language interfaces for databases based on chatgpt and langchain. In *Proc. 42nd Int. Conf. on Conceptual Modeling–Posters&Demos, Lisbon, Portugal*.

Open Ai (2024). Gpt-4 is openai's most advanced system, producing safer and more useful responses. `https://openai.com/index/gpt-4/`.

Papadopoulos, A. V., Versluis, L., Bauer, A., Herbst, N., Kistowski, J. v., Ali-Eldin, A., Abad, C. L., Amaral, J. N., Tůma, P., and Iosup, A. (2021). Methodological principles for reproducible performance evaluation in cloud computing. *IEEE Transactions on Software Engineering*, 47(8):1528–1543.

Patel, P., Choukse, E., Zhang, C., Goiri, I. n., Warrier, B., Mahalingam, N., and Bianchini, R. (2024). Characterizing power management opportunities for llms in the cloud. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS '24, page 207–222, New York, NY, USA. Association for Computing Machinery.

Pericherla, S. S. (2023). Cloud computing threats, vulnerabilities and countermeasures: A state-of-the-art. *The ISC International Journal of Information Security*, 15:1–58.

Ren, J., Fu, D., Shi, C., Huang, Z., Zhu, W., and Liu, Y. (2023). Research on cloud computing technology graph analysis. pages 84–91. Institute of Electrical and Electronics Engineers Inc.

Saini, H., Upadhyaya, A., and Khandelwal, M. K. (2019). Benefits of cloud computing for business enterprises: A review. *SSRN Electronic Journal*.

Salesforce (2024). The history of salesforce. https://www.salesforce.com/news/stories/the-history-of-salesforce/.

Sinha, S. (2023). Azure advisor. https://www.codingninjas.com/studio/library/azure-advisor.

Soygazi, F. and Oguz, D. (2023). An analysis of large language models and langchain in mathematics education. In *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence*, pages 92–97.

Steve Chipman (2024). A brief history of google workspace. https://www.lexnetcg.com/blog/google-workspace/brief-history/.

Surbiryala, J. and Rong, C. (2019). Cloud computing: History and overview. *Proceedings - 2019 3rd IEEE International Conference on Cloud and Fog Computing Technologies and Applications, Cloud Summit 2019*, pages 1–7.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89(425):208–218.

Topsakal, O. and Akinci, T. C. (2023). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences*, volume 1, pages 1050–1056.

Tozzi, C. (2023). The benefits and limitations of google cloud recommender. https://www.techtarget.com/searchcloudcomputing/tip/The-benefits-and-limitations-of-Google-Cloud-Recommender.

Wang, Q., Li, J., Wang, S., Xing, Q., Niu, R., Kong, H., Li, R., Long, G., Chang, Y., and Zhang, C. (2024a). Towards next-generation llm-based recommender systems: A survey and beyond. *arXiv preprint arXiv:2410.19744*.

Wang, Z., Zhu, Y., Chen, M., Liu, M., and Qin, W. (2024b). Llm connection graphs for global feature extraction in point cloud analysis. *Applied Science and Biotechnology Journal for Advanced Research*, 3(4):10–16.

A. Conceptos de nube extendidos

A.1. Breve historia de la computación en la nube

Antes de adentrarnos en la historia de la computación en la nube, es esencial comprender los términos que dan forma a este concepto revolucionario. La palabra “nube” denota la provisión de servicios a través de Internet. Por otro lado, “computación” abarca el cálculo, procesamiento y ejecución de tareas y operaciones proporcionadas por un ordenador. Es crucial destacar que la visión de la computación en la nube no es un desarrollo reciente. En 1961, el Instituto Tecnológico de Massachusetts (MIT) profetizó que el poder de cómputo podría, en algún momento, organizarse como un servicio público, análogo al suministro de agua, energía eléctrica o servicios telefónicos. Esta predicción visionaria es la piedra angular sobre la cual se ha construido la noción de la computación en la nube, visualizando proveedores especializados que democratizarían el acceso al poder de cómputo, haciéndolo tan accesible como otros servicios fundamentales para la sociedad. Este concepto sentó las bases para la evolución y la materialización de la computación en la nube tal como la conocemos hoy en día (Garfinkel and Abelson, 1999).

En la década de los 70s, la empresa IBM puso en marcha una innovación crucial al introducir el concepto de Máquinas Virtuales. Este enfoque revolucionario permitía la creación de ambientes virtualizados, donde múltiples sistemas operativos podían compartir el mismo hardware y simular una infraestructura de computadoras interconectadas. Esta implementación brindó a los usuarios y empresas una libertad sin precedentes, ya que les permitía focalizarse en el diseño y la interconexión de redes, liberándolos de la preocupación por la compleja implementación y gestión de la infraestructura de hardware subyacente. La introducción de las máquinas virtuales marcó un hito significativo en la evolución de la informática, allanando el camino para futuros desarrollos en el campo de la virtualización y sentando las bases para la computación en la nube tal como la conocemos hoy (Agarwal et al., 2016).

A finales de la década de los 90, la compañía Salesforce desencadenó una revolución al permitir a los usuarios emplear software empresarial sin la necesidad de instalarlo en cada computadora cliente. Este hito marcó la introducción temprana del concepto de Software como Servicio (SaaS - Software as a Service), donde la capacidad de cómputo se proporciona a través de internet (Salesforce, 2024). Poco después, en el año 2004, Google lanzó Google Apps, consolidando aún más el uso de servicios de software a través de la web. Estos servicios no solo estaban disponibles para empresas, sino también para usuarios finales

(Hart-Davis, 2021) (Steve Chipman, 2024). Este cambio hacia la entrega de software como servicio a través de la nube allanó el camino para transformaciones radicales en la forma en que las organizaciones y los individuos acceden y utilizan aplicaciones informáticas.
En el año 2006, Amazon tomó la decisión estratégica de introducir el servicio de Infraestructura como Servicio (IaaS) mediante su producto Amazon Elastic Compute Cloud (EC2) . Este servicio innovador proporciona la capacidad de crear y virtualizar máquinas virtuales dentro de la infraestructura propia de Amazon. Este lanzamiento pionero marcó un hito significativo en la evolución de la computación en la nube (AWS, 2024). Posteriormente, empresas líderes como IBM, Microsoft, Alibaba, Oracle, Google, entre otros, siguieron el ejemplo de Amazon presentando sus propios servicios que abarcan desde infraestructura y plataformas hasta contenedores, funciones y software como servicio (SaaS). Este impulso hacia la diversificación de servicios en la nube pública ha enriquecido continuamente el mercado, proporcionando a los usuarios una amplia gama de opciones y soluciones para satisfacer diversas necesidades. Desde entonces, la competencia y la innovación en el ámbito de la nube pública han florecido, brindando a las organizaciones acceso a herramientas cada vez más avanzadas y eficientes.

A.2. Tipos de computación en la nube

En la sección anterior, se proporcionó un resumen conciso de los eventos clave en la evolución de la computación en la nube. No obstante, es imperativo profundizar en la comprensión de este fenómeno, reconociendo que la computación en la nube se manifiesta en distintas formas, destacando principalmente los modelos de nube pública, privada e híbrida. Estos enfoques ofrecen variedad y flexibilidad para satisfacer las necesidades específicas de las organizaciones y usuarios en el complejo paisaje tecnológico actual.

A.2.1. Nube Pública

La **nube pública** es un modelo de servicio en la computación en la nube en el cual los recursos informáticos, como servidores, almacenamiento y redes, son proporcionados por un proveedor de servicios externo a través de Internet. Estos recursos son compartidos por múltiples clientes, lo que permite la escalabilidad y el acceso bajo demanda a una variedad de servicios y aplicaciones. Algunas características clave de la nube pública incluyen:

- Los servicios en la nube pública están disponibles desde cualquier lugar con conexión a Internet, lo que brinda una flexibilidad significativa a los usuarios.
- Los clientes pagan solo por los recursos que consumen, lo que permite un modelo de costos más eficiente y escalable. Esto contrasta con los modelos tradicionales en los que se adquieren y mantienen recursos físicos independientemente del uso real.

- Los proveedores de nube pública ofrecen la capacidad de escalar recursos de manera dinámica según las necesidades del usuario. Esto permite gestionar picos de demanda sin la necesidad de invertir en infraestructuras fijas.
- La responsabilidad de gestionar y mantener la infraestructura recae en el proveedor de servicios, liberando a los usuarios de la carga operativa asociada con la administración de servidores y hardware.

La nube pública se utiliza comúnmente para una variedad de aplicaciones, desde el alojamiento de sitios web y el almacenamiento de datos hasta la implementación de aplicaciones empresariales y servicios de inteligencia artificial. Aunque ofrece una flexibilidad y escalabilidad notables, los usuarios deben considerar cuestiones de seguridad y privacidad al confiar en recursos gestionados externamente.

A.2.2. Nube Privada

La **nube privada** se destaca como un modelo de implementación que fusiona control, seguridad y personalización a medida. A diferencia de su contraparte pública, la nube privada ofrece una infraestructura dedicada a una única entidad, proporcionando un terreno fértil para la innovación y la adaptabilidad. Aquí desglosamos las características clave que definen este paradigma tecnológico:

- La nube privada concede un control exclusivo a la organización, permitiendo una adaptabilidad excepcional. Este nivel de autonomía es esencial para satisfacer las demandas únicas de las organizaciones modernas.
- Mantener los recursos en el ámbito de la organización aborda inquietudes fundamentales de seguridad y privacidad. Esto se vuelve imperativo en industrias altamente reguladas, donde la confidencialidad de los datos es esencial.
- La nube privada permite a las organizaciones adaptar la infraestructura según sus necesidades. Desde integrar aplicaciones heredadas hasta configurar medidas de seguridad personalizadas, la personalización sin restricciones es una ventaja distintiva.
- La implementación de una nube privada puede ocurrir tanto en las instalaciones de la organización como a través de un proveedor externo. Este despliegue estratégico ofrece opciones que se alinean con los objetivos y recursos de la organización.
- A pesar de los costos iniciales potencialmente más altos, la nube privada puede resultar económicamente efectiva a largo plazo para organizaciones con cargas de trabajo predecibles y consistentes, brindando un control que justifica la inversión inicial.
- Sectores como finanzas, salud y gobierno, que manejan datos críticos, son ejemplos de áreas donde la nube privada se destaca como la elección preferida debido a sus características exclusivas.

En la intersección de control absoluto y tecnología avanzada, la nube privada emerge como un faro para organizaciones que buscan un manejo riguroso de sus recursos en la nube, desbloqueando un potencial ilimitado y un terreno fértil para la innovación a medida.

A.2.3. Nube Híbrida

En el paisaje tecnológico actual, la nube híbrida se presenta como una solución innovadora, fusionando lo mejor de la nube pública y privada para satisfacer las necesidades dinámicas y complejas de las empresas modernas. Aquí se desglosa de manera detallada los aspectos clave que definen este paradigma tecnológico:

- La nube híbrida representa una combinación de entornos de nube pública y privada, permitiendo a las organizaciones integrar de manera fluida aplicaciones, datos y cargas de trabajo entre estos dos dominios. Esta integración sin fisuras facilita la optimización de recursos y el flujo de trabajo eficiente.
- La capacidad de combinar recursos locales con la escalabilidad dinámica de la nube pública brinda una flexibilidad excepcional. Las organizaciones pueden adaptarse rápidamente a las fluctuaciones de demanda, aprovechando recursos adicionales de la nube pública cuando sea necesario y manteniendo cargas de trabajo críticas en entornos privados.
- A través de herramientas de gestión unificada, las empresas pueden administrar y supervisar sus entornos en la nube híbrida de manera centralizada. Esto simplifica la administración de recursos distribuidos y mejora la eficiencia operativa.
- La nube híbrida aborda las preocupaciones de seguridad y cumplimiento al permitir que datos sensibles permanezcan en entornos privados, mientras que las cargas de trabajo menos críticas se despliegan en la nube pública. Esto brinda un equilibrio entre la seguridad necesaria y la flexibilidad requerida para la innovación.
- Al tener redundancia en múltiples entornos, la nube híbrida mejora la resiliencia y garantiza la continuidad del negocio. En caso de fallos en un entorno, las aplicaciones y datos pueden migrar automáticamente a otro, garantizando una operación sin interrupciones.
- La nube híbrida permite una estrategia de migración gradual, donde las organizaciones pueden trasladar cargas de trabajo específicas a la nube pública según sus propios plazos y requisitos. Esto facilita la transición sin interrupciones y minimiza los riesgos asociados.
- Sectores como el financiero, la salud y la fabricación, que tienen requisitos diversos y a menudo críticos, son ejemplos donde la nube híbrida se convierte en una solución estratégica para equilibrar eficacia, innovación y cumplimiento normativo.

La nube híbrida representa un enfoque equilibrado y adaptable para las organizaciones que buscan la máxima flexibilidad y eficiencia. Al conectar lo mejor de ambos mundos, la nube híbrida se erige como un puente tecnológico que impulsa la transformación digital y responde a los desafíos de un entorno empresarial en constante evolución.

B. Análisis de selección para servicios de fundación de Google Cloud

B.1. Fundación de Google Cloud: Organización

¿Candidato para ser analizado?: No

Descripción: Un recurso de organización en Google Cloud representa su empresa y actúa como el nodo de nivel superior en su jerarquía. Para crear su organización, debe configurar un servicio de identidad de Google y asociarlo con su dominio. Al completar este proceso, se crea automáticamente un recurso de organización.

Justificación: Debido a que una organización en Google Cloud representa un conjunto amplio de proyectos, carpetas y recursos, junto con la inclusión de productos adicionales como Cloud Identity y Google Workspace, la automatización de su análisis se vuelve altamente compleja. Por lo tanto, siguiendo el paso 5 del flujo de decisión (figura **1-1**), en el que se considera que no es viable la implementación de la automatización del análisis, este componente se marca como fuera del alcance del presente trabajo.

B.2. Fundación de Google Cloud: Usuarios y grupos

¿Candidato para ser analizado?: No

Descripción: Un grupo es una colección nombrada de Cuentas de Google y cuentas de servicio. Cada grupo tiene una dirección de correo electrónico única, como gcp-billing-admins@ejemplo.com. Los grupos se crean para gestionar usuarios y aplicar roles de IAM a gran escala.

A continuación se recomiendan los siguientes grupos para ayudar a administrar las funciones clave de la organización y completar el proceso de configuración de Google Cloud.

Tabla **B-1**.: Grupos administrativos recomendados para Google Cloud 20 de noviembre de 2024

Grupo	Descripción
gcp-organization-admins	Administrar todos los recursos de la organización. Asignar este rol solo a los usuarios más confiables.
gcp-billing-admins	Configurar cuentas de facturación y monitorear el uso.

Grupo	Descripción
gcp-network-admins	Crear redes de Virtual Private Cloud, subredes y reglas de firewall.
gcp-hybrid-connectivity-admins	Crear dispositivos de red como instancias de Cloud VPN y Cloud Router.
gcp-logging-admins	Utilizar todas las funciones de Cloud Logging.
gcp-logging-viewers	Acceso de solo lectura a un subconjunto de registros.
gcp-monitoring-admins	Administradores de monitoreo tienen acceso a todas las funciones de Cloud Monitoring.
gcp-security-admins	Establecer y gestionar políticas de seguridad para toda la organización, incluyendo la gestión de acceso y las políticas de restricciones organizativas. Para más información, consulte el blueprint de fundaciones empresariales de Google Cloud para planificar su infraestructura de seguridad.
gcp-developers	Diseñar, codificar y probar aplicaciones.
gcp-devops	Crear o gestionar pipelines completos que soporten la integración y entrega continua, monitoreo y aprovisionamiento de sistemas.

Fuente: Elaboración propia y adaptación de (GCP Recommender, 2023)

Justificación: Aunque Google Cloud recomienda ciertos grupos administrativos para tareas específicas, como se observa en la tabla **B-1**, no existe una regla estricta y universal que defina los usuarios y grupos recomendados para el uso de Google Cloud, ya que estos dependen de la estructura y necesidades de cada organización.

Por lo tanto, de acuerdo con el flujo de decisión presentado en la figura **1-1**, la tarea 3 establece que no hay una documentación definitiva sobre las mejores prácticas relacionadas con la gestión de usuarios y grupos. En consecuencia, el análisis de esta característica fundacional queda fuera del alcance de este trabajo.

B.3. Fundación de Google Cloud: Acceso de administrador

¿Candidato para ser analizado?: **No**

Descripción: Para otorgar el acceso adecuado a cada grupo de administradores creado en la tarea de Usuarios y grupos, revise los roles predeterminados asignados a cada grupo. Puede agregar o eliminar roles para personalizar el acceso de cada grupo según sea necesario.

Justificación: Tal como ocurre en el caso anterior, aunque Google Cloud recomienda la creación de grupos de administración con accesos y roles sugeridos, esto depende de cada organización. Por lo tanto, no existe una documentación única aplicable a todos los casos. En consecuencia, la tarea 3 del flujo de decisión **1-1** establece que esta característica fundacional debe considerarse fuera del alcance del análisis.

B.4. Fundación de Google Cloud: Facturación

¿Candidato para ser analizado?: No

Descripción: Las cuentas de Cloud Billing están vinculadas a uno o más proyectos de Google Cloud y se utilizan para pagar los recursos que utiliza, como máquinas virtuales, redes y almacenamiento.

- **Autogestionada (o en línea)**: Regístrese en línea utilizando una tarjeta de crédito o débito. Se recomienda esta opción si es una pequeña empresa o un individuo. Cuando se registra en línea para una cuenta de facturación, su cuenta se configura automáticamente como una cuenta autogestionada.
- **Facturada (o fuera de línea)**: Si ya tiene una cuenta de facturación autogestionada, puede ser elegible para solicitar una facturación por factura si su negocio cumple con los requisitos de elegibilidad.

Justificación: Si bien existe documentación que incluye buenas prácticas para la gestión de cuentas de facturación, este es un elemento financiero que varía según cada organización. Por lo tanto, en virtud de la tarea 5 del flujo de decisión, no es viable automatizar este análisis, ya que no es posible obtener información sobre las cuentas de facturación a través de la API para determinar si cumplen o no con las buenas prácticas. Además, las cuentas de facturación no son parte de los elementos de decisión arquitectónica, ya que se consideran un aspecto meramente administrativo.

B.5. Fundación de Google Cloud: Jerarquía de acceso

¿Candidato para ser analizado?: No

Descripción: Crear una estructura de carpetas y proyectos ayuda a gestionar los recursos de Google Cloud y a asignar acceso según el funcionamiento de la organización. Por ejemplo, se puede organizar y proporcionar acceso a los recursos en función de la colección única de regiones geográficas, estructuras de subsidiarias o marcos de responsabilidad de la organización.
La jerarquía de recursos permite establecer límites y compartir recursos en toda la organización para tareas comunes. Esta jerarquía puede configurarse utilizando una de las siguientes estructuras iniciales, de acuerdo con la estructura organizativa:

- Entorno simple orientado a entornos:
  - Aislar entornos como No producción y Producción.
  - Implementar políticas, requisitos regulatorios y controles de acceso específicos en cada carpeta de entorno. Adecuado para empresas pequeñas con entornos centralizados.
- Entorno simple orientado a equipos:

  - Aislar equipos como Desarrollo y Control de Calidad (QA).
  - Aislar el acceso a los recursos mediante carpetas de entorno secundarias bajo cada carpeta de equipo.
  - Adecuado para pequeñas empresas con equipos autónomos.
- Orientado a Entornos:
  - Priorizar el aislamiento de entornos como No producción y Producción.
  - Bajo cada carpeta de entorno, aislar las unidades de negocio.
  - Bajo cada unidad de negocio, aislar los equipos.
  - Ideal para grandes empresas con entornos centralizados.
- Orientado a unidades de negocio:
  - Priorizar el aislamiento de unidades de negocio como Recursos Humanos e Ingeniería para asegurar que los usuarios solo accedan a los recursos y datos que necesitan.
  - Bajo cada unidad de negocio, aislar los equipos.
  - Bajo cada equipo, aislar los entornos.
  - Adecuado para grandes empresas con equipos autónomos.

Cada configuración incluye una carpeta común para proyectos que contienen recursos compartidos, como los proyectos de registro y monitoreo.
Justificación: La organización de las carpetas y la jerarquía de recursos depende de cada entidad y de los permisos asignados a los equipos de administración tecnológica de la misma. Debido a esta variabilidad, no es posible definir una documentación única que establezca las mejores prácticas para analizar dentro de la configuración de la jerarquía de recursos. Por lo tanto, con base en la tarea 3 del flujo de decisión **1-1**, se concluye que, ante la ausencia de una documentación estándar de mejores prácticas, no es viable automatizar el análisis de este módulo.

B.6. Fundación de Google Cloud: Registro Centralizado de Logs

¿Candidato para ser analizado?: **No**
Descripción: Cloud Logging facilita el almacenamiento, búsqueda, análisis, monitoreo y la creación de alertas sobre los datos de registro y eventos generados en Google Cloud. Además, permite recopilar y procesar registros provenientes de aplicaciones, recursos locales y otras nubes. Se recomienda utilizar Cloud Logging para consolidar los registros en un único bucket de registros, centralizando así la gestión y

el análisis de logs.

Justificación: El registro centralizado de logs no cuenta con una documentación exhaustiva de buenas prácticas más allá de las políticas de auditoría de logs. Además, la configuración de la centralización de registros varía según cada organización, lo que impide definir un análisis estándar de buenas prácticas aplicable a todos los casos. Por lo tanto, de acuerdo con las tareas 3 y 5 del flujo de decisión **1-1**, se concluye que este módulo queda fuera del alcance del presente trabajo.

B.7. Fundación de Google Cloud: Seguridad

¿Candidato para ser analizado?: No

Descripción: Para centralizar los servicios de reporte de vulnerabilidades y amenazas, se recomienda habilitar el *Security Command Center* (Centro de Comando de Seguridad). Esta herramienta ayuda a fortalecer la postura de seguridad y mitigar riesgos.

Justificación: Esta tarea de fundación se centra en la habilitación del servicio Security Command Center de Google Cloud, al menos en su versión más básica. Esta herramienta se encarga de realizar el escaneo de vulnerabilidades, riesgos y otras configuraciones de seguridad de los componentes configurados en Google Cloud. Por lo tanto, la tarea 5 del flujo de decisión **1-1** marca esta tarea como fuera del alcance, dado que ya existe un producto dedicado a la recomendación de mejores prácticas de seguridad. No es necesario realizar un aporte adicional para este módulo fundacional de Google Cloud.

B.8. Fundación de Google Cloud: Redes de VPC

¿Candidato para ser analizado?: Si

Descripción: Una red de Virtual Private Cloud (VPC) es una versión virtual de una red física que se implementa dentro de la red de producción de Google. Una red VPC es un recurso global compuesto por subredes regionales.

Las redes VPC proporcionan capacidades de red a los recursos de Google Cloud, tales como instancias de máquinas virtuales de Compute Engine, contenedores de GKE, e instancias del entorno flexible de App Engine.

Shared VPC conecta recursos de varios proyectos a una red VPC común, lo que les permite comunicarse entre sí utilizando las direcciones IP internas de la red. Al usar Shared VPC, se designa un proyecto host

y se adjuntan uno o más proyectos de servicio. Las redes de Virtual Private Cloud en el proyecto host se denominan redes de Shared VPC. Es posible usar un proyecto host para gestionar centralmente lo siguiente:

- Rutas
- Reglas de cortafuegos
- Conexiones VPN
- Subredes

Un proyecto de servicio es cualquier proyecto que esté vinculado a un proyecto host. Es posible compartir subredes, incluidas las de rangos secundarios, entre los proyectos host y de servicio.
Cada red de Shared VPC contiene subredes públicas y privadas:

- La subred pública puede ser utilizada por instancias orientadas a internet para conectividad externa.
- La subred privada puede ser utilizada por instancias internas que no tienen direcciones IP públicas asignadas.

Justificación: El módulo de redes VPC se presenta como un candidato ideal para el análisis de buenas prácticas de infraestructura, dado que los servicios que operan en la nube suelen requerir integración con otros componentes de software ubicados tanto en Google Cloud como en redes on-premises, lo que lo convierte en un producto de alto uso (tarea 6 del flujo **1-1**). Además, debido a su carácter crítico, existe una vasta documentación que establece buenas prácticas específicas para su configuración (tarea 3), y muchas de estas no están completamente cubiertas por el servicio de Google Recommender (GCP Recommender, 2023) (tarea 4).
Otro factor clave que favorece la inclusión de este módulo en el análisis es que las configuraciones actuales de las redes VPC se pueden obtener de manera programática, lo que facilita la evaluación y ajuste de dichas configuraciones en función de las buenas prácticas recomendadas. Por lo tanto, este módulo será parte del análisis y configuración como parte de este trabajo.

B.9. Fundación de Google Cloud: Conectividad híbrida

¿Candidato para ser analizado?: Si
Descripción: Este proceso crea una VPN de Alta Disponibilidad (HA VPN), que es una solución de alta disponibilidad que se puede implementar rápidamente para transmitir datos a través de internet público.

Este tipo de conexiones permiten establecer enlaces de baja latencia y alta disponibilidad entre redes VPC y redes **on-premises** u otras redes en la nube. Para configurar este tipo de conexiones, se deben establecer los siguientes componentes:

- Puerta de enlace HA VPN de Google Cloud: Un recurso regional que cuenta con dos interfaces, cada una con su propia dirección IP. Es posible especificar el tipo de pila IP, lo que determinará si se admite tráfico IPv6 en la conexión.
- Puerta de enlace VPN del par (peer VPN gateway): La puerta de enlace en la red del par, a la cual se conecta la puerta de enlace HA VPN de Google Cloud. Se deben ingresar las direcciones IP externas que la puerta de enlace del par utiliza para conectarse a Google Cloud.
- Cloud Router: Utiliza el Protocolo de Puerta de Enlace Fronteriza (BGP) para intercambiar rutas dinámicamente entre las redes VPC y las redes del par. Se asigna un Número de Sistema Autónomo (ASN) como identificador para el Cloud Router, y se especifica el ASN que utiliza el enrutador del par.
- Túneles VPN: Conectan la puerta de enlace de Google Cloud con la puerta de enlace del par. Es necesario especificar el protocolo de intercambio de claves de Internet (IKE) que se utilizará para establecer el túnel. Se puede ingresar una clave IKE previamente generada o generar y copiar una nueva.

Justificación: La conectividad de Google Cloud hacia otras nubes o sistemas *on-premises* mediante VPN o Interconnect es un caso de uso bastante común, ya que estos sistemas se comunican de manera segura a través de estos canales. Por ello, Google Cloud ha identificado buenas prácticas para crear estos canales de comunicación híbrida (tarea 3), las cuales no están completamente cubiertas por Google Recommender (tarea 4). Además, es posible obtener la configuración de manera programática (tarea 5) para analizar estas configuraciones, las cuales son ampliamente utilizadas (tarea 6). Por lo tanto, resulta relevante realizar un análisis sobre este módulo fundacional dentro del presente trabajo (ver flujo **1-1**).

B.10. Fundación de Google Cloud: Monitoreo

¿Candidato para ser analizado?: No

Descripción: Cloud Monitoring recopila métricas, eventos y metadatos de los servicios de Google Cloud, monitores sintéticos, instrumentación de aplicaciones y otros componentes comunes de aplicaciones. Este servicio se configura automáticamente para los proyectos de Google Cloud, facilitando la supervisión continua de los recursos y servicios sin necesidad de una configuración manual adicional.

Justificación: Los tableros de monitoreo proporcionan métricas, estadísticas y datos informativos sobre el estado de los servicios configurados en Google Cloud, así como en sistemas on-premises donde se

encuentre instalado el agente de Google Cloud. La instrumentación de estas métricas y la creación de los tableros pueden variar considerablemente entre los usuarios, lo que implica que no existe una documentación de buenas prácticas que estandarice las configuraciones de monitoreo. Por lo tanto, de acuerdo con la tarea 3 del flujo de decisión presentado en la figura **1-1**, este módulo fundacional queda fuera del alcance del análisis.

B.11. Fundación de Google Cloud: Soporte

¿Candidato para ser analizado?: No

Descripción: Un plan de soporte premium proporciona asistencia crítica para el negocio, lo que permite resolver rápidamente problemas con la ayuda de expertos de Google Cloud.

Justificación: Este módulo fundacional de Google Cloud sugiere a los usuarios seleccionar un plan de soporte que se adecue a sus necesidades. No se dispone de documentación específica ni se requiere un análisis de buenas prácticas para este módulo, ya que se trata simplemente de la adquisición de un plan de soporte destinado a atender las demandas de asistencia de los clientes de Google Cloud.

Una vez completado el análisis de los módulos de fundación de Google Cloud que son candidatos para ser evaluados en este trabajo, en los capítulos posteriores se presentará una descripción de la herramienta que facilitará dicho análisis. Esta herramienta se aplicará a los módulos seleccionados en esta sección, los cuales fueron elegidos siguiendo el flujo de selección de módulos detallado en la sección 1.2.1.

C. Glosario

C.1. Modelo *"Hub-and-spoke"*

El modelo *"hub-and-spoke"*, también conocido como *"concentrador y radios"*, es una arquitectura de red que se utiliza para conectar múltiples redes (los *"radios"* o *"spokes"*) a una red central (el "concentrador" o "hub"). Esta arquitectura se asemeja a una rueda de bicicleta, donde el concentrador es el centro y los radios son las conexiones que se extienden hacia afuera. En el contexto de la nube, el hub suele ser una VPC (Virtual Private Cloud) que actúa como un punto central de conectividad y control. Los spokes son otras VPCs que se conectan al hub a través de conexiones seguras como VPNs o peering de VPC (Google Cloud Architecture Center, 2024). El tráfico entre los spokes se enruta a través del hub, lo que permite la implementación de políticas de seguridad centralizadas y la inspección del tráfico (Google Cloud Architecture Center, 2024). La administración de la red se centraliza en el hub, lo que facilita la configuración, el monitoreo y el mantenimiento. Así, se optimiza el uso de recursos al compartir servicios comunes en el hub, como firewalls, gateways de VPN y servidores DNS (Google Cloud Architecture Center, 2024).

C.2. Proyectos *"Spoke"*

En la arquitectura *"hub-and-spoke"*, los proyectos spoke son aquellos que se implementan en las VPCs que actúan como radios". Estos proyectos suelen contener aplicaciones o servicios específicos que se benefician del aislamiento y la seguridad que proporciona la arquitectura *"hub-and-spoke"* (Google Cloud Architecture Center, 2024).

C.3. Formato *"Markdown"*

Markdown es un lenguaje de marcado ligero que se utiliza para dar formato a texto de forma sencilla. Utiliza caracteres de texto plano para indicar diferentes elementos de formato, como encabezados, listas, enlaces, imágenes y código. Es ampliamente usado con el fin de documentar y versionar el código fuente de ciertas aplicaciones (Markdown Guide, 2024).

C.4. *"Slots"* en BigQuery

En Google BigQuery, los slots son unidades de procesamiento que se utilizan para ejecutar consultas. Cada slot representa una CPU virtual que se asigna a una consulta para su procesamiento (GCP BigQuery, 2023). BigQuery ofrece dos modelos de precios:

- **Precios bajo demanda**: BigQuery calcula automáticamente la cantidad de slots necesarios para una consulta y cobra en función del volumen de datos procesados.
- **Precios basados en la capacidad**: Los usuarios reservan una cantidad específica de slots y pagan una tarifa fija por la capacidad reservada, independientemente del volumen de datos procesados.