Costos de Implementación YoCampo Prototipo

Análisis de Costos y Estrategias de Optimización para la Implementación de YoCampo en Azure

Autores: Victor Manuel Mondragon Maca – Laboratorio de Datos SNUIRA

Versión:1.0

Fecha: 15 de julio de 2024

Unidad de Planificación Rural Agropecuaria (UPRA)
Calle 28 Nº 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.
+57(601) 552 9820, 245 7307
www.upra.gov.co

Unidad de Planificación Rural Agropecuaria (UPRA)
Calle 28 Nº 13-22, Torre C, piso 3. Edif. Palma Real. Bogotá, Colombia.
+57(601) 552 9820, 245 7307
www.upra.gov.co

Resumen

Este documento proporciona un análisis de los costos asociados a la implementación de YoCampo Prototipo experimental, una solución basada en Azure que utiliza múltiples servicios avanzados para ofrecer funcionalidades robustas y escalables. El objetivo principal es identificar y desglosar los costos operativos, así como proponer estrategias de optimización para una gestión eficiente del presupuesto. En el documento se presenta la arquitectura y los diferentes servicios como Azure Cognitive Search, Cognitive Services, Virtual Machines, Azure App Service, Storage, Microsoft Defender for Cloud, Load Balancer y otros servicios adicionales. Presenta los costos generales y específicos por producto dentro de cada servicio, proporcionando una visión granular de los gastos.

Tabla de contenido

Índice de tablas

Índice de figuras

Lista de siglas y abreviaturas

Aunap	Autoridad Nacional de Acuicultura y Pesca
UPRA	Unidad de Planificación Rural Agropecuaria

Glosario

Azure Cognitive Search: Un servicio de búsqueda en la nube con capacidades de inteligencia artificial, utilizado para realizar búsquedas avanzadas y recuperación de información.

Cognitive Services: Conjunto de servicios de inteligencia artificial que incluye procesamiento de lenguaje natural, visión artificial y servicios de habla.

Virtual Machines (VM): Instancias de servidores virtuales en la nube, altamente configurables y escalables.

Azure App Service: Plataforma totalmente gestionada para construir, desplegar y escalar aplicaciones web.

Azure Storage: Servicio de almacenamiento en la nube altamente escalable y duradero.

Microsoft Defender for Cloud: Servicio de seguridad que proporciona gestión unificada de la seguridad y protección contra amenazas para las cargas de trabajo en la nube.

Azure Load Balancer: Servicio que distribuye el tráfico entrante a varios servicios en la nube, asegurando alta disponibilidad y redundancia.

Azure Cosmos DB: Base de datos NoSQL globalmente distribuida, diseñada para ofrecer escalabilidad y disponibilidad altas.

Blob Storage: Servicio de almacenamiento de objetos de Azure, optimizado para almacenar grandes cantidades de datos no estructurados.

Azure OpenAI: Servicio que proporciona acceso a modelos avanzados de inteligencia artificial de OpenAI, como GPT-4, a través de la infraestructura de Azure.

Azure AI: Conjunto de servicios y herramientas de inteligencia artificial ofrecidos por Azure, diseñados para ayudar a los desarrolladores a construir aplicaciones inteligentes.

Budget (Presupuesto): Límite financiero establecido para controlar y monitorear los gastos en Azure.

Forecast (Proyección): Estimación de costos futuros basada en el uso actual y tendencias pasadas.

Actual (Real): Costos efectivamente incurridos en un periodo específico.

Cost Management (Gestión de Costos): Herramienta de Azure que permite monitorizar, asignar y optimizar los costos asociados a los servicios utilizados.

Scoping (Alcance): Proceso de asignar y definir un presupuesto a un ámbito específico dentro de Azure, como suscripciones, grupos de recursos o servicios específicos.

Resource (Recurso): Cualquier instancia de servicio o componente de Azure utilizado en la implementación de YoCampo.

Resource Group (Grupo de Recursos): Contenedor lógico en Azure que agrupa recursos relacionados para facilitar su administración.

Subscription (Suscripción): Unidad de facturación y administración en Azure que organiza recursos y servicios.

Cost Allocation (Asignación de Costos): Proceso de distribuir los costos a diferentes proyectos, departamentos o unidades de negocio dentro de una organización.

Cost Optimization (Optimización de Costos): Estrategias y acciones para reducir los costos operativos mientras se mantiene o mejora el rendimiento y la eficiencia de los servicios en la nube.

YoCampo: Proyecto que utiliza múltiples servicios de Azure para ofrecer una solución robusta y escalable en el ámbito agropecuario.

Retrieval-Augmented Generation (RAG): Técnica que combina la recuperación de documentos relevantes con la generación de texto, mejorando la precisión y relevancia de las respuestas generadas por modelos de lenguaje.

Region (Región): Ubicación geográfica de los centros de datos de Azure donde se alojan los recursos.

Scope (Ámbito): El rango o alcance al que se aplica un presupuesto o política en Azure.

Introducción

Este documento tiene como finalidad proporcionar un análisis detallado de los costos asociados a la implementación de la solución YoCampo. Basada en la arquitectura RAG (Retrieval-Augmented Generation) y utilizando servicios avanzados de Azure, YoCampo es una aplicación que ofrece asistencia inteligente y personalizada a investigadores, técnicos, extensionistas y productores agropecuarios. Este análisis se enfoca en los costos generales y particulares de los componentes de la arquitectura de solución, permitiendo una visión clara y precisa de los recursos económicos necesarios para el prototipo experimental de YoCampo. La información se basa en los componentes implementados en la arquitectura Cloud de Azure Microsoft.

Objetivos

- Identificar y desglosar los costos asociados a cada uno de los componentes principales de la arquitectura de solución RAG implementada en Azure.
- Evaluar la distribución de costos entre diferentes servicios de Azure, ubicaciones geográficas y recursos específicos.
- Proporcionar una base de referencia para la gestión financiera del proyecto, facilitando la toma de decisiones informadas sobre la optimización y escalabilidad de la solución.
- Pautas para el monitorear y proyectar los costos mensuales, asegurando que el presupuesto asignado se utilice de manera eficiente y efectiva.

Alcance

Este documento abarca los siguientes aspectos:

- Descripción de la Arquitectura de Solución: Un resumen de los componentes clave de la arquitectura RAG utilizada en YoCampo, basada en la información proporcionada en el archivo CSV y la imagen de costos mensuales.
- Análisis Detallado de Costos por Componente: Desglose de los costos de cada servicio de Azure empleado en la solución, incluyendo Azure Key Vault, Azure App Service Plan, Azure AI Hub, Azure Cosmos DB, y reglas de alerta de Azure Smart Detector.
- Distribución de Costos por Servicio: Evaluación de los costos asociados a diferentes servicios de Azure utilizados en la implementación de YoCampo, identificando las áreas de mayor inversión.
- Distribución de Costos por Ubicación: Análisis de cómo se distribuyen los costos entre diferentes ubicaciones geográficas, proporcionando una perspectiva sobre la eficiencia regional de los recursos.
- Distribución de Costos por Recurso: Identificación de los recursos específicos dentro de cada servicio que contribuyen a los costos generales, permitiendo un enfoque más granular en la optimización de gastos.
- Proyecciones y Tendencias de Costos Mensuales: Evaluación de los costos acumulados y previstos para el mes de julio de 2024, ayudando a anticipar futuros gastos y ajustando el presupuesto en consecuencia.

1. Arquitectura Cloud YoCampo

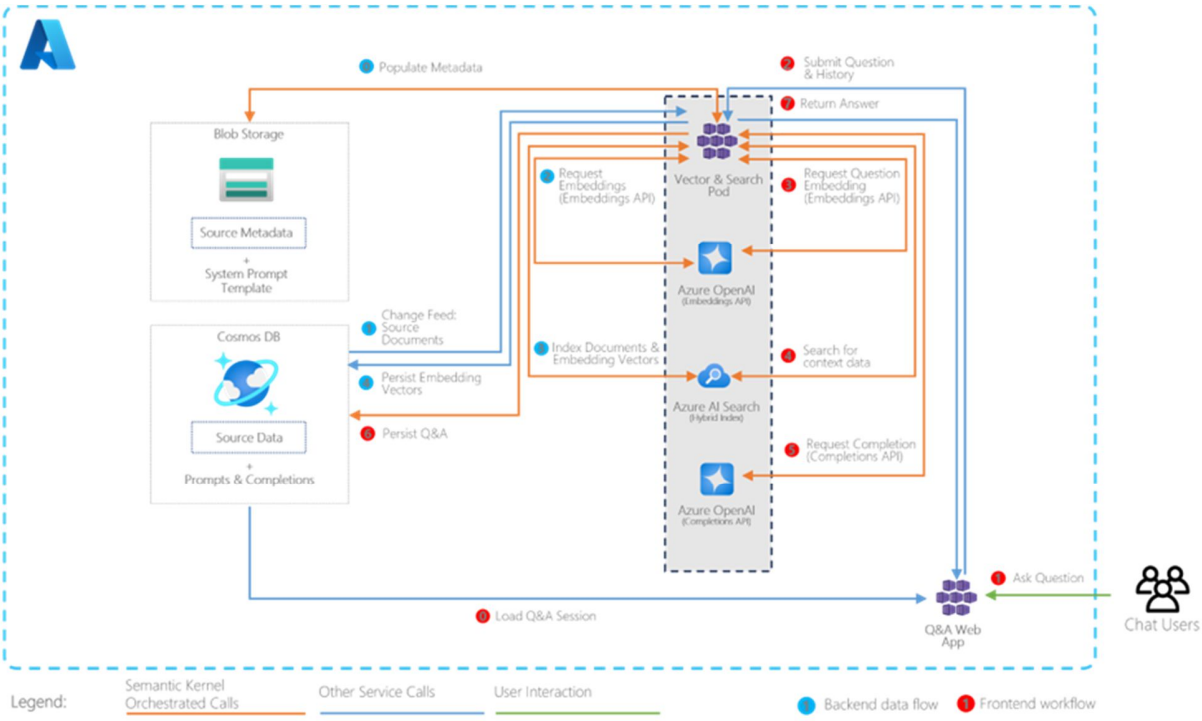
El proyecto YoCampo, utiliza inteligencia artificial generativa para asistir a los actores del sector agropecuario con información actualizada y personalizada. La arquitectura RAG facilita la integración de estos componentes mediante la combinación de tecnologías de recuperación de información y generación de lenguaje natural. Al implementar una solución RAG, YoCampo puede ofrecer respuestas precisas y personalizadas a consultas específicas, de esta forma mejorando el acceso a la información para el beneficio de actores del Sistema Nacional de Innovación Agropecuaria (SNIA).

A continuación, se describe la arquitectura de solución implementada en el prototipo experimental YoCampo aliñada a la estimación de costos para su replicación y proyecciones de operación.

1.1. Arquitectura de Solución

La arquitectura de solución representada en la Figura 1 RAG (Retrieval-Augmented Generation) se compone de los siguientes elementos principales:

Figura 1. Ejemplo de mapa de contexto general



Fuente: [1](2024).

- Blob Storage: Almacena metadatos y plantillas de prompts del sistema.
- Cosmos DB: Guarda datos fuente y vectores de incrustaciones, así como preguntas y respuestas (Q&A).
- Azure OpenAI: Proporciona API para incrustaciones y completaciones, que se utilizan para generar vectores y respuestas.
- Azure AI Search: Realiza búsquedas contextuales en los datos indexados.
- Vector & Search Pod: Maneja las solicitudes de incrustaciones y búsquedas vectoriales.

1.2. Flujo de Datos

- Carga de Sesión Q&A: Inicializa la sesión de preguntas y respuestas.
- Envío de Preguntas: Los usuarios envían preguntas y el historial al sistema.

- Solicitudes de Incrustaciones: Las preguntas se procesan para generar vectores.
- Búsqueda de Datos Contextuales: Utiliza los vectores para buscar información relevante.
- Solicitud de Competición: Genera respuestas usando la API de completar.
- Persistencia de Q&A: Las respuestas se almacenan en Cosmos DB [2].
- Retorno de Respuestas: Las respuestas generadas se devuelven a los usuarios.

En resumen, la arquitectura de solución RAG [3] proporciona la infraestructura necesaria para que YoCampo pueda gestionar y procesar grandes volúmenes de datos, generar respuestas en tiempo real, y ofrecer un servicio eficiente y efectivo a los investigadores, extensionistas y productores agropecuarios en Colombia.

2. Análisis de Costos

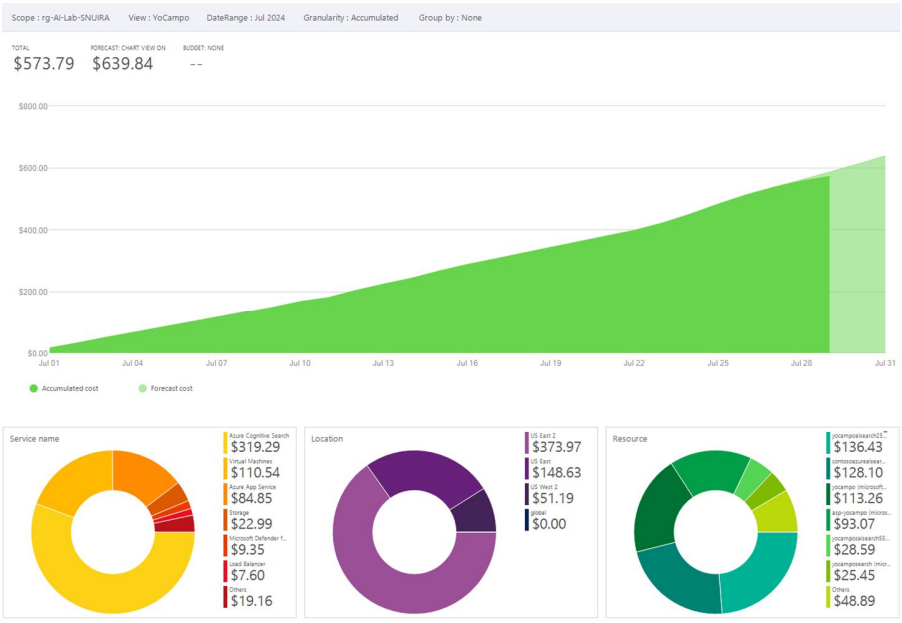
A continuación, se realiza una descripción de los costos de implementación general y específicos de YoCampo como piloto experimental. Los valores se encuentran descritos para el periodo piloto de implementación final del prototipo comprendido entre el 1 al 26 de julio de 2024.

2.1. Análisis Costos Generales

La Figura 2 describe el total de Costos acumulados y previstos en la implementación del piloto, con el siguiente resumen:

- Costo total acumulado hasta la fecha del presente informe: $573.79
- Costo previsto para el final del mes: $639.84

Figura 2. Costos generales YoCampo



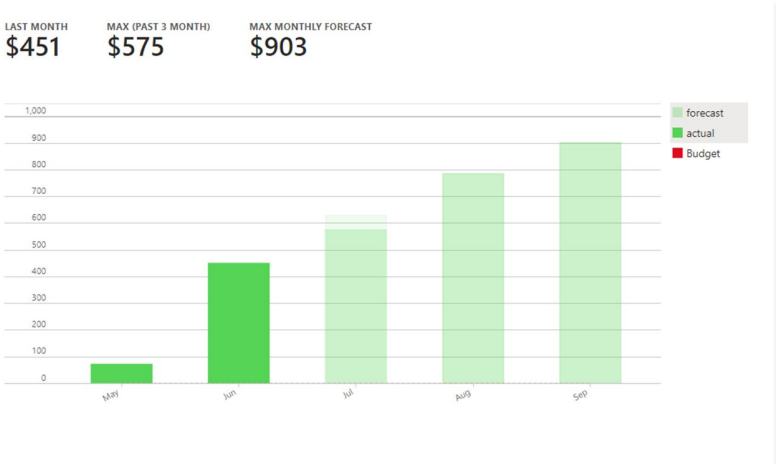
Fuente: Propia Azure (2024).

2.2. Proyección de costos YoCampo

El análisis de la Figura 3 muestra que el proyecto YoCampo ha experimentado un aumento constante en los costos mensuales, especialmente notable en junio. Consecuencia de la implementación de los servicios cognitivos. Las proyecciones indican un aumento en los próximos meses, con un costo máximo mensual esperado de $903. Lo cual representa costos de elementos de pago por uso. Se proyecta Implementar estrategias de optimización puede ayudar a gestionar estos costos de manera más eficiente, evitando sobrepasar el presupuesto establecido.

Este análisis proporciona una visión clara del comportamiento de los costos y las proyecciones futuras para YoCampo, permitiendo una gestión proactiva y la toma de decisiones informadas.

Figura 3 Proyección de costo YoCampo prototipo.



Fuente: Propia Azure (2024).

2.2.1. **Descripción de variables**

A continuación, se describen las variables observadas en la Figura 3:

- Last Month ($451): Este valor representa el costo total incurrido en el último mes. Indica cuánto se gastó en los servicios de Azure durante ese período específico.
- Max (Past 3 Month) ($575): Este valor representa el costo máximo mensual incurrido en los últimos tres meses. Muestra el mes con el mayor gasto dentro de ese período.
- Max Monthly Forecast ($903): Este valor representa la proyección máxima de costos mensuales futuros. Indica el gasto esperado más alto en los próximos meses basado en el uso actual y las tendencias.
- Forecast (verde claro): La barra verde clara representa la proyección de costos para los próximos meses. Esta es una estimación basada en el uso actual y las tendencias pasadas.
- Actual (verde oscuro): La barra verde oscuro representa los costos reales incurridos. Esta barra muestra los gastos efectivos ya realizados.
- Budget (rojo): La línea roja (no visible en la imagen) representaría el presupuesto establecido para los gastos. Ayuda a comparar los costos proyectados y reales contra el presupuesto planificado.

2.2.2. **Análisis de los costos proyectados**

La imagen proporciona una visión clara de los costos y proyecciones para el proyecto YoCampo en Azure, distribuidos en un gráfico de barras que abarca varios meses.

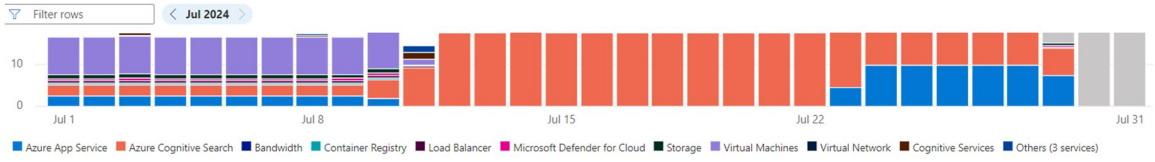
- Costo del Último Mes (Junio): El costo del último mes es de $451, lo que muestra un gasto significativo comparado con el mes anterior (mayo), donde los costos fueron muy bajos.

- Costo Máximo de los Últimos Tres Meses (junio): El mes con el mayor gasto en los últimos tres meses es junio, con un total de $575. Esto sugiere un aumento en el uso de los servicios de Azure durante este mes.
- Proyección Máxima Mensual ($903): La proyección para el futuro mes (septiembre) muestra un costo esperado de $903, lo que indica un aumento considerable en el uso de los servicios de Azure. Esta proyección puede estar basada en tendencias de uso crecientes y en la escalabilidad del proyecto.
- Comparación de Costos Actuales y Proyectados:
    - Julio: Los costos reales para julio son de aproximadamente $573, con una proyección ligeramente superior, indicando una tendencia de aumento en el uso de servicios.
    - Agosto y Septiembre: Los costos proyectados para estos meses son constantes en alrededor de $900, lo que sugiere una expectativa de uso sostenido y posiblemente incremento de servicios.
    -

2.3. Costos de servicio

La implementación de YoCampo se basa en varios servicios de Azure, cada uno desempeñando un rol crucial en la arquitectura de la solución. A continuación, se presenta un análisis detallado de los costos asociados a cada servicio, su descripción y cómo se integran en la arquitectura de YoCampo.

Figura 4 Costos por servicio



Fuente: Propia Azure(2024).

2.3.1. Azure Cognitive Search

- Descripción del Servicio: Azure Cognitive Search [4] es un servicio de búsqueda en la nube con capacidades de inteligencia artificial. Permite a los desarrolladores agregar una potente funcionalidad de búsqueda en sus aplicaciones web y móviles.
- Rol en YoCampo: Utilizado para realizar búsquedas vectoriales y contextuales en los datos almacenados. Es fundamental para la funcionalidad de recuperación de información y generación de respuestas en tiempo real.
- Costos Detallados:
    - Standard S1 Unit: Unidad estándar para capacidades de búsqueda avanzada y rendimiento. Costo: $293.84 USD
    - Basic Unit: Unidad básica para funcionalidades esenciales de búsqueda. Costo: $25.45 USD.
- Costo Asociado: $319.29 USD/ mes

2.3.2. Virtual Machines

- Análisis de Costos: Este es el componente de mayor costo en la arquitectura actual. La optimización de las consultas y la gestión eficiente de los índices pueden ayudar a reducir este costo.
- Descripción del Servicio: Las Máquinas Virtuales de Azure permiten la creación de instancias de servidores virtuales en la nube. Son altamente configurables y escalables.
- Rol en YoCampo: Utilizadas para ejecutar aplicaciones y servicios que requieren un entorno controlado y configuraciones específicas.
-
- Costos Detallados [5]:
    - D2 v2/DS2 v2:
        - Descripción: Máquinas virtuales con 2 núcleos y 7 GB de RAM.
        - Costo: $86.02 USD
    - D4 v3/D4s v3:

    - Descripción: Máquinas virtuales con 4 núcleos y 16 GB de RAM.
    - Costo: $9.69 USD
  - D12 v2/DS12 v2:
    - Descripción: Máquinas virtuales con 12 núcleos y 56 GB de RAM.
    - Costo: $4.67 US

- Costo Asociado: $110.54 USD /mes

El uso de instancias reservadas y la automatización del apagado de VM fuera de horas pico pueden contribuir a reducir costos

2.3.3. **Azure App Service**

- Descripción del Servicio: Azure App Service es una plataforma totalmente gestionada para construir, desplegar y escalar aplicaciones web.
- Rol en YoCampo: Facilita la implementación y el escalado de la aplicación web de YoCampo (https://yocampo.azurewebsites.net/), proporcionando una plataforma robusta y escalable.
- Costo detallado:
  - App Service Plan (Linux): Descripción: Plan de servicio para aplicaciones web basadas en Linux.
- Costo Asociado: $84.85 USD /mes

Asegurarse de que se utiliza el plan de servicio adecuado y optimizar el escalado automático puede ayudar a gestionar y reducir costos

2.3.4. **Storage**

- **Descripción del Servicio**: Azure Storage ofrece almacenamiento altamente escalable y duradero para datos en la nube.
- **Rol en YoCampo**: Utilizado para almacenar metadatos, datos fuente y resultados de consultas.
- App Service Plan (Linux): Descripción: Plan de servicio para aplicaciones web basadas en Linux.

- **Costo Asociado**: $22.99 USD

Implementar políticas de ciclo de vida y eliminar datos no necesarios puede ayudar a minimizar costos de almacenamiento.

2.3.5. Microsoft Defender for Cloud

- Descripción del Servicio: Proporciona gestión unificada de la seguridad y protección contra amenazas para las cargas de trabajo en la nube.
- Rol en YoCampo: Asegura que todos los componentes de la solución estén protegidos contra amenazas y vulnerabilidades.
- Costo Asociado: $9.35 USD

2.3.6. Load Balancer

- Descripción del Servicio: Azure Load Balancer distribuye el tráfico entrante a varios servicios en la nube, asegurando alta disponibilidad y redundancia.
- Rol en YoCampo: Distribuye las solicitudes de usuarios entre múltiples instancias de servicios para asegurar una alta disponibilidad y equilibrio de carga.
- Costo Asociado: $7.60 USD

2.3.7. Cognitive Services

- Descripción del Servicio: Azure Cognitive Services proporciona una amplia gama de capacidades de inteligencia artificial, incluyendo procesamiento de lenguaje natural, visión artificial y servicios de habla [6].
- Rol en YoCampo: Utilizado para procesamiento de lenguaje natural y generación de respuestas a través de modelos avanzados como GPT-4 [7].
- Costos Detallados:

- embedding-ada-regional Tokens:Tokens utilizados para la generación de incrustaciones utilizando el modelo Ada.Costo: $4.07 USD
- gpt-4o-Input-regional Tokens: Tokens de entrada utilizados para el procesamiento de texto utilizando GPT-4 [7]. Costo: $3.05 USD
- Costo Total Asociado: $7.54 USD

Estos costos mensuales detallados permiten entender mejor los gastos asociados a cada servicio en la arquitectura de YoCampo. Al conocer estos costos, se pueden implementar estrategias de optimización y ajustes necesarios para gestionar eficientemente el presupuesto y asegurar una implementación sostenible de la solución.

2.4. Costos por Recurso

A continuación, se relaciona los recursos de la arquitectura con su respectiva descripción:

Recurso: yocampoaisearch250863064198

- ResourceType: Search service
- ResourceLocation: US East
- ServiceName: Azure Cognitive Search
- Producto: Standard S1 Unit
- Costo: $136.79 USD

yocampo

- ResourceType: microsoft.machinelearningservices/workspaces
- ResourceLocation: US East 2, US East
- ServiceName: Virtual Machines
- Producto: D2 v2/DS2 v2
- Costo: $86.02 USD

Recurso: yocampo

- ResourceType: microsoft.machinelearningservices/workspaces
- ResourceLocation: US East 2, US East
- ServiceName: Storage
- Producto: P6 LRS Disk
- Costo: $9.42 USD

Recurso: yocampo

- ResourceType: microsoft.machinelearningservices/workspaces
- ResourceLocation: US East 2, US East
- ServiceName: Load Balancer
- Producto: Standard Included LB Rules and Outbound Rules
- Costo: $7.08 USD

Recurso: yocampo

- ResourceType: microsoft.machinelearningservices/workspaces
- ResourceLocation: US East 2, US East
- ServiceName: Virtual Machines
- Producto: NC48ads_A100_v4
- Costo: $3.16 USD

Recurso: yocampoaisearch586520058698

- ResourceType: Search service
- ResourceLocation: US East 2
- ServiceName: Azure Cognitive Search
- Producto: Standard S1 Unit
- Costo: $128.10 USD
- Análisis Detallado por Recurso

Recurso: yocampoaisearch250863064198

- Descripción del Producto: Unidad estándar para capacidades de búsqueda avanzada y rendimiento.

- Costo en US East: $136.79 USD

Recurso: yocampo

- Virtual Machines (D2 v2/DS2 v2):
- Descripción: Máquinas virtuales con 2 núcleos y 7 GB de RAM.
- Costo: $86.02 USD
- Storage (P6 LRS Disk):
- Descripción: Almacenamiento de propósito general, versión 2, para múltiples tipos de datos.
- Costo: $9.42 USD

Load Balancer (Standard Included LB Rules and Outbound Rules):

- Descripción: Balanceador de carga básico para distribuir el tráfico de manera eficiente.
- Costo: $7.08 USD
- Virtual Machines (NC48ads_A100_v4):
- Descripción: Máquinas virtuales con configuraciones avanzadas para cargas de trabajo intensivas.
- Costo: $3.16 USD

Recurso: yocampoaisearch586520058698

- Descripción del Producto: Unidad estándar para capacidades de búsqueda avanzada y rendimiento.
- Costo en US East 2: $128.10 USD

El análisis detallado de costos por recurso muestra que Azure Cognitive Search es el componente más costoso de la implementación de YoCampo, seguido por los costos de las máquinas virtuales y otros servicios asociados. Implementar estrategias de optimización específicas para estos servicios, como ajustar las

configuraciones según la demanda y revisar los planes de servicio, puede ayudar a reducir los costos operativos totales.

Un mayor detalle se puede observar en el Anexo 1. Tabla de costos por recurso detallada.

2.5. Análisis Detallado de Costos por Región

Este análisis proporciona una visión resumida de los costos asociados a la implementación de YoCampo en diferentes regiones según Microsoft Azure. La información está basada en los datos proporcionados por la arquitectura Azure y se presenta en formato de tabla y gráfico para una comprensión más clara.

Tabla 1. Tabla de distribución de costos por región Azure

Región	Costo Total (USD)
US East	$142.41
US East 2	$134.30
US East 2, US East	$113.26
US East 2, global	$0,006
US East, US East 2	$0.45
US West 2	$25.87
global, US East	$0.0013
global, US East 2	$0.0014

Fuente: UPRA (2024).

El análisis de costos por región muestra que la mayoría de los costos están concentrados en las regiones de US East y US East 2. Estos costos se deben principalmente a servicios de Azure Cognitive Search y máquinas virtuales.

Implementar estrategias de optimización específicas para estos servicios y regiones, como ajustar las configuraciones según la demanda y revisar los planes de servicio, puede ayudar a reducir los costos operativos totales. Además, considerar la migración de algunos recursos a regiones con menores costos puede resultar en ahorros significativo.

2.6. Costos No incluidos.

En particular los costos asociados a los servicios cognitivas OPENAI fueron obtenidos con el siguiente detalle:

2.6.1. Tokens de Entrada y Salida

Tokens de Entrada: Los costos se generan por la cantidad de tokens que se envían al modelo para su procesamiento.

Tokens de Salida: Los costos también se aplican a los tokens generados por el modelo en respuesta a una solicitud.

2.6.2. Modelos Utilizados

Modelos GPT-4-o: Los costos pueden variar dependiendo de la versión del modelo (por ejemplo, GPT-4, GPT-4-turbo, etc.).

Modelos Embedding: Los costos asociados a la generación de incrustaciones (embeddings) para la búsqueda y recuperación de información.

Descripción: Tokens utilizados para la generación de incrustaciones utilizando el modelo Ada.

Costo: $4.07 USD (ejemplo basado en datos proporcionados)

gpt-4o-Input-regional Tokens

Descripción: Tokens de entrada utilizados para el procesamiento de texto utilizando GPT-4. Costo: $3.05 USD (ejemplo basado en datos proporcionados)

gpt-4o-Output-regional Tokens

Descripción: Tokens de salida utilizados para la generación de texto utilizando GPT-4-o. Costo: $0.42 USD (ejemplo basado en datos proporcionados)

Factores que Influyen en los Costos

- Uso del Modelo: El tipo de modelo utilizado (por ejemplo, GPT-4 versus un modelo de menor capacidad) afectará los costos.
- Cantidad de Tokens Procesados: Mayor cantidad de tokens procesados resultará en costos más altos.
- Región: La ubicación geográfica donde se procesan los datos puede influir en los costos debido a variaciones en precios regionales.
- Frecuencia de Uso: El uso continuo o intensivo de los modelos resultará en costos acumulativos más altos.

Sin embargo, para realizar un calculo de los costos por utilización de Tokens OPEN AI se proyecta prueba piloto con usuarios finales por un periodo de 2 semas.

3. Estrategias de Optimización de Costos

La implementación de una solución basada en múltiples servicios de Azure implicar costos significativos. A continuación, se presentan varias estrategias para optimizar estos costos y garantizar una utilización eficiente de los recursos:

3.1. Dimensionamiento Correcto de Recursos

- **Monitorización y Ajuste Dinámico**: Utiliza Azure Monitor y Application Insights para supervisar el uso de los recursos y ajustar dinámicamente el tamaño de las instancias de acuerdo con la carga de trabajo. Reduce el tamaño durante los periodos de baja demanda.
- **Escalado Automático**: Implementa políticas de escalado automático para ajustar los recursos en función de la demanda en tiempo real. Esto es especialmente útil para servicios como Azure App Service y Azure Virtual Machines.

3.2. Selección de Niveles de Servicio Apropiados

- **Planes de Servicio Adecuados**: Elige el plan de servicio adecuado que equilibre costos y rendimiento. Por ejemplo, utiliza el Plan Básico para entornos de desarrollo y pruebas, y el Plan Estándar o Premium para producción. En el caso de YoCampo fueron utilizadas en el despliegue Plan Básico.
- **Revisar Niveles de Servicio**: Evalúa periódicamente los niveles de servicio contratados y ajusta según las necesidades actuales del proyecto.

3.3. Optimización de Almacenamiento

- **Gestión de Almacenamiento en Azure Blob Storage y Cosmos DB**: Implementa políticas de ciclo de vida para mover datos a niveles de almacenamiento más económicos cuando ya no se necesiten accesos

frecuentes. En el caso de YoCampo se utilizando contenedores con capacidad mínima implementada de 5 TB

- **Compresión y Eliminación de Datos**: Comprime los datos y elimina los que no sean necesarios para reducir el consumo de almacenamiento.

3.4. Uso de Servicios de Azure más Económicos

- **Azure Reserved Instances**: Aprovechar las instancias reservadas de Azure para servicios como Virtual Machines y Cosmos DB, que pueden ofrecer descuentos significativos para compromisos a largo plazo.
- **Azure Spot Instances**: Utiliza instancias spot para cargas de trabajo que no sean críticas y puedan tolerar interrupciones.

3.5. Optimización de Bases de Datos

- **Ajuste de Capacidad de Cosmos DB**: Configura la capacidad de Cosmos DB (Request Units - RUs) según las necesidades reales y ajusta periódicamente.
- **Consolidación de Bases de Datos**: Revisa la posibilidad de consolidar bases de datos para reducir el número de instancias necesarias.

3.6. Optimización de Redes y Tráfico

- **Uso de Azure Traffic Manager**: Implementa Azure Traffic Manager para enrutar el tráfico de manera eficiente y reducir la latencia, lo cual puede disminuir costos asociados al uso de redes.
- **Minimización del Tráfico Saliente**: Reduce el tráfico de datos salientes mediante el uso de cachés y compresión de datos.
- **Suscripciones y Descuentos**: Aprovecha las suscripciones de nivel empresarial y descuentos ofrecidos por Azure para organizaciones.

La optimización de costos es un proceso continuo que requiere monitoreo, ajuste y la implementación de estrategias específicas para asegurar que los recursos

se utilizan de manera eficiente y económica. Adoptar estas estrategias puede ayudar a reducir significativamente los costos de la implementación de YoCampo, permitiendo una gestión más efectiva del presupuesto y recursos disponibles.

3.7. Estrategias de Optimización por recurso

- Optimización de Azure Cognitive Search
    - Revisión de Consultas y Índices: Asegúrate de que las consultas sean eficientes y que los índices estén bien configurados para evitar costos innecesarios.
    - Reducción de Capacidad Durante Periodos de Baja Demanda: Ajusta la capacidad del servicio según la demanda para reducir costos.
- Optimización de Máquinas Virtuales
    - Uso de Instancias Reservadas: Considera el uso de instancias reservadas si las cargas de trabajo son predecibles, lo que puede ofrecer descuentos significativos.
    - Automatización del Apagado de Máquinas: Implementa scripts para apagar máquinas virtuales fuera de horas pico.
- Optimización del Servicio de Aplicaciones (App Service)
    - Selección del Plan de Servicio Adecuado: Asegúrate de que el plan seleccionado se ajuste a las necesidades de rendimiento sin exceder lo necesario.
    - Escalado Automático: Utiliza el escalado automático para ajustar los recursos según la demanda real.
- Optimización del Almacenamiento
    - Compresión y Eliminación de Datos No Necesarios: Implementa políticas de ciclo de vida para mover datos a niveles de almacenamiento más económicos y eliminar datos no necesarios.
- Revisión de Otros Servicios y Recurso

- Revisión Periódica de Recursos: Revisa y elimina recursos no utilizados o infrautilizados.
- Monitoreo y Alertas: Configura monitoreos y alertas para recibir notificaciones cuando los costos superen ciertos umbrales.

el análisis detallado de los costos de la implementación de YoCampo en Azure ha identificado varias áreas clave donde se pueden realizar optimizaciones. Al enfocarse en servicios específicos como Azure Cognitive Search, Virtual Machines y App Service, se pueden implementar estrategias de reducción de costos efectivas. Además, una revisión periódica y ajustes basados en la demanda real contribuirán significativamente a una gestión financiera más eficiente del proyecto.

4. Conclusiones

- Costos Principales y Estrategias de Optimización: Azure Cognitive Search y Virtual Machines representan los costos más significativos en la implementación de YoCampo. Es esencial optimizar estos servicios ajustando configuraciones según la demanda y revisando los planes de servicio. Estas acciones son cruciales para reducir gastos operativos y mantener la eficiencia financiera del proyecto.
- Distribución de Costos por Recurso: Recursos específicos, como yocampoaisearch250863064198 y contosoazureaisearch586520058698, concentran una parte considerable de los gastos. Monitorear y gestionar estos recursos de manera proactiva es fundamental para identificar oportunidades de ahorro y mejorar la eficiencia operativa.
- Análisis de Costos por Región: Las regiones de US East y US East 2 son las más costosas debido al uso intensivo de servicios clave. Evaluar la migración de algunos recursos a regiones con menores costos y optimizar la gestión de recursos en estas áreas puede ayudar a disminuir significativamente los gastos operativos.
- Proyecciones de Costos y Monitoreo Continuo: Las proyecciones indican un aumento significativo en los costos mensuales futuros, alcanzando hasta $903. Es importante en la implementación un monitoreo continuo y alertas para gestionar los gastos de manera proactiva, asegurando que los costos se mantengan dentro de los límites presupuestarios y evitando sobrecostos.
- Se debe realizar pruebas con usuarios finales para calcular costos de acceso OPENAI.

5. Referencias

[1] Microsoft Azure AI, «github - Vector-Search-AI-Assistant,» github , Marzo 2024. [En línea]. Available: https://github.com/Azure/Vector-Search-AI-Assistant/tree/cognitive-search-vector. [Último acceso: mayo 2024].

[2] Microsoft-architecture, «Architecting applications on Azure,» 20 06 2023. [En línea]. Available: https://learn.microsoft.com/en-us/azure/architecture/.

[3] Mondragon, V.M y García-Díaz, V, «Adaptive contents for interactive TV guided by machine learning based on predictive sentiment analysis of data,» *Springer.*

[4] Microsoft, «Azure Machine Learning prompt flow,» 01 mayo 2024. [En línea]. Available: https://learn.microsoft.com/en-us/azure/machine-learning/prompt-flow/overview-what-is-prompt-flow?view=azureml-api-2.

[5] microsoft, «Arquitectura de análisis moderno con Azure Databricks,» 23 Sep 2023. [En línea]. Available: https://learn.microsoft.com/es-es/azure/architecture/solution-ideas/articles/azure-databricks-modern-analytics-architecture.

[6] ScienceDirect IBM, «A Method for Implementation of Machine Learning Solutions for Predictive Maintenance in Small and Medium Sized Enterprises,» *ELSEVIER,* pp. https://pdf.sciencedirectassets.com/282173/1-s2.0-S2212827120X00102/1-s2.0-S2212827120306223/, 2020.

[7] B. Peng y C. Li, «Instruction Tuning with GPT-4,» *arXiv,* 2023.

Anexos

Anexo 1. Tabla de costos por recurso detallada.

Tabla 2 Tabla detalles costos Jul 01-31, 2024

ServiceName	ServiceTier	Meter	CostUSD
Azure App Service	Azure App Service Free Plan	F1 App	USD 0,000
Azure App Service	Azure App Service Free Plan - Linux	F1 App	USD 0,000
Azure App Service	Azure App Service Std Plan-Linux	S1 App	USD 23,899
Azure App Service	Azure App Service Std Plan-Linux	S3 App	USD 61,766
Azure Cognitive Search	Azure AI Search	Basic Unit	USD 25,454
Azure Cognitive Search	Azure AI Search	Standard S1 Unit	USD 294,202
Azure Cosmos DB	Azure Cosmos DB	Data Stored	USD 0,000
Azure Cosmos DB	Azure Cosmos DB serverless	1M RUs	USD 0,000
Bandwidth	Bd Inter-Rgn	Inter Continent Data Transfer Out - NAM or EU To Any	USD 0,000
Bandwidth	Bd Inter-Rgn	Intra Continent Data Transfer Out	USD 0,065
Bandwidth	Rtn Pref: MGN	Standard Data Transfer In	USD 0,000
Bandwidth	Rtn Pref: MGN	Standard Data Transfer Out	USD 0,839
Cognitive Services	OpenAI	embedding-ada-regional Tokens	USD 1,951
Cognitive Services	OpenAI	embedding-ada-regional Tokens	USD 2,122
Cognitive Services	OpenAI	gpt-4o-Input-regional Tokens	USD 1,153
Cognitive Services	OpenAI	gpt-4o-Input-regional Tokens	USD 1,900
Cognitive Services	OpenAI	gpt-4o-Output-regional Tokens	USD 0,247
Cognitive Services	OpenAI	gpt-4o-Output-regional Tokens	USD 0,169
Container Registry	Container Registry	Basic Registry Unit	USD 6,913

Key Vault	Key Vault	Operations	USD 0,005
Load Balancer	Load Balancer	Standard Data Processed	USD 0,254
Load Balancer	Load Balancer	Standard Included LB Rules and Outbound Rules	USD 7,347
Microsoft Defender for Cloud	Defender for Azure Cosmos DB	Standard 100 RU/s	USD 0,000
Microsoft Defender for Cloud	Microsoft Def for App Service	Standard Node	USD 8,271
Microsoft Defender for Cloud	Microsoft Def for Key Vault	Standard Transactions	USD 0,004
Microsoft Defender for Cloud	Microsoft Def for Servers	Standard P2 Node	USD 0,775
Microsoft Defender for Cloud	Microsoft Def for Storage	Standard Transactions	USD 0,344
Storage	Azure Data Lake Storage Gen2 Flat Namespace	Hot Other Operations	USD 0,000
Storage	Blob Storage	All Other Operations	USD 0,002
Storage	Blob Storage	All Other Operations	USD 0,017
Storage	Blob Storage	Hot LRS Data Stored	USD 0,386
Storage	Blob Storage	Hot LRS Data Stored	USD 0,084
Storage	Blob Storage	Hot LRS Write Operations	USD 0,076
Storage	Blob Storage	Hot LRS Write Operations	USD 0,109
Storage	Blob Storage	Hot Read Operations	USD 0,003
Storage	Blob Storage	Hot Read Operations	USD 0,017
Storage	Blob Storage	LRS List and Create Container Operations	USD 0,004
Storage	Blob Storage	LRS List and Create Container Operations	USD 0,360
Storage	Files	Delete Operations	USD 0,000
Storage	Files	LRS Data Stored	USD 0,176

Storage	Files	LRS Write Operations	USD 0,008
Storage	Files	List Operations	USD 0,001
Storage	Files	Protocol Operations	USD 0,000
Storage	Files	Read Operations	USD 0,007
Storage	Premium Block Blob	Premium LRS All Other Operations	USD 0,000
Storage	Premium Block Blob	Premium LRS Data Stored	USD 0,826
Storage	Premium Block Blob	Premium LRS List and Create Container Operations	USD 0,003
Storage	Premium Block Blob	Premium LRS Read Operations	USD 0,001
Storage	Premium Block Blob	Premium LRS Write Operations	USD 0,005
Storage	Premium SSD Managed Disks	P10 LRS Disk	USD 10,713
Storage	Premium SSD Managed Disks	P10 LRS Disk	USD 0,257
Storage	Premium SSD Managed Disks	P10 LRS Disk	USD 0,493
Storage	Premium SSD Managed Disks	P6 LRS Disk	USD 9,416
Storage	Premium SSD Managed Disks	P6 LRS Disk	USD 0,015
Storage	Queues v2	Class 2 Operations	USD 0,000
Storage	Queues v2	Class 2 Operations	USD 0,000
Storage	Queues v2	LRS Class 1 Operations	USD 0,000
Storage	Queues v2	LRS Class 1 Operations	USD 0,000
Storage	Standard SSD Managed Disks	E10 LRS Disk	USD 0,028
Storage	Standard SSD Managed Disks	E4 LRS Disk Operations	USD 0,038
Storage	Tables	Batch Write Operations	USD 0,001
Storage	Tables	LRS Data Stored	USD 0,001
Storage	Tables	Read Operations	USD 0,008
Storage	Tables	Scan Operations	USD 0,000
Storage	Tables	Write Operations	USD 0,000

Virtual Machines	BS Series Windows VM	B4ms	USD 0,232
Virtual Machines	Dv2 Series VM	D12 v2/DS12 v2	USD 4,675
Virtual Machines	Dv2 Series VM	D2 v2/DS2 v2	USD 86,020
Virtual Machines	Dv3 Series VM	D4 v3/D4s v3	USD 2,013
Virtual Machines	Dv3 Series Windows VM	D4 v3/D4s v3	USD 9,692
Virtual Machines	Edsv4 Series VM	E4ds v4	USD 2,419
Virtual Machines	Ev3 Series VM	E4 v3/E4s v3	USD 2,001
Virtual Machines	Ev3 Series VM	E4 v3/E4s v3	USD 0,326
Virtual Machines	NCads A100 v4 Series Linux	NC48ads_A100_v4	USD 3,165
Virtual Machines Licenses	RHEL	4 vCPU VM License	USD 0,027
Virtual Network	IP Address	Standard IPv4 Static Public IP	USD 3,801