SSMAS – IA – Casos de uso

Presenter’s name here

Introducción

- SSMAS, una empresa española de tecnología publicitaria (AdTech) que se ha consolidado como un actor fundamental en el mercado de la monetización de contenidos digitales. El análisis revela que la posición de liderazgo de SSMAS se fundamenta en su condición de pionera como la primera compañía en España en obtener la certificación de Google Certified Publishing Partner (GCPP). Esta acreditación no es meramente un título, sino el pilar de su modelo de negocio, otorgándole una ventaja competitiva significativa en términos de acceso a tecnología, credibilidad en el mercado y conocimiento profundo del ecosistema publicitario de Google.

- El modelo de negocio de SSMAS se centra en la optimización programática de ingresos para editores y creadores de contenido digital. Su propuesta de valor se materializa en un incremento promedio de ingresos del 37% para sus clientes durante los primeros meses de 2024, gestionando un volumen superior a los 7.000 millones de impresiones publicitarias mensuales. Este rendimiento se apoya en una cartera de servicios y tecnología propia, destacando su solución "SSM. CODES", que simplifica la compleja implementación de tecnologías como Header Bidding, plataformas de gestión de consentimiento (CMPs) y Open Bidding de Google.

Entendimiento de la necesidad

SSMAS indica a INGRAM que esta interesado en el uso de Inteligencia Artificial utilizando el hiperacelerador AWS, lo que proporcionará a SSMAS una mejora competitiva versus competencia.

- SSMAS dispone de una plataforma de datos desplegada en Vercel, que en la actualidad gestiona mas de 30 millones de registros al día.
- SSMAS expone mediante API (elastic share) de consumo de datos, aunque previsiblemente para los casos de uso planteados, sea necesario un desarrollo que permita un consumo mas eficiente dada la volumetría de datos actuales.

Dentro de los potenciales usos de AWS en el campo de la IA, SSMAS comunico a INGRAM que esta interesado en los siguientes casos de uso:

- **Yield Predictivo:** La solución consiste en desarrollar un modelo de machine learning avanzado que predice el Coste Por Mil (CPM) esperado para cada oportunidad de anuncio individual.
- **Utilización de bedrock y sus capacidades agenticas**, que permita automatizar tareas repetitivas dentro de los sus procesos actuales actuales. El objetivo ambicioso es automatizar por completo el ciclo de monitorización, diagnóstico y resolución de problemas operativos 24/7.

Visión Ingram

La plataforma propuesta se adhiere a cuatro principios fundamentales, diseñados para garantizar la agilidad, la eficiencia y la seguridad:

1.Serverless-First: Priorizar el uso de servicios gestionados y sin servidor para minimizar la carga operativa, reducir los costes de gestión de infraestructuras y permitir que el equipo de ingeniería se centre en la creación de valor empresarial en lugar de en el mantenimiento de servidores.

2.En Tiempo Real por Diseño: La publicidad programática opera en milisegundos. La arquitectura debe estar diseñada desde cero para la ingesta, el procesamiento y la acción de baja latencia, permitiendo a SSMAS competir en la velocidad del mercado.

3.Modular y Evolutiva: La arquitectura debe ser un conjunto de componentes modulares y débilmente acoplados. Esto permite una implementación por fases, donde cada nuevo caso de uso se construye sobre la base existente sin requerir una reingeniería masiva, asegurando que la plataforma pueda evolucionar con las necesidades del negocio.

4.Segura por Defecto: La confianza es la moneda del ecosistema AdTech. La arquitectura debe incorporar las mejores prácticas de seguridad y gobernanza de datos desde el principio, utilizando herramientas para el cifrado, la gestión de identidades y el control de acceso de grano fino para proteger los datos sensibles de editores y usuarios.

Visión Ingram

Plan

Fase I

- El caso de uso se implanta con una cuenta AWS del cliente.
- Se realiza en un entorno Poc controlado y aislado **sin conexión a la infraestructura actual del cliente**.
- El cliente proporciona una muestra de datos amplia para ser procesada.
- Los datos se procesan (feature engineering).
- Se entrena el modelo y se analiza el comportamiento.
- Si todos los resultados son adecuados, el MVP esta listo para operativizar.

Fase II

- Se operativiza el MVP aplicando buenas practicas:
- Si realiza análisis de necesidades para operativizar el mvp:
  - Se planteará el uso de buenas practicas como landing zone, seguridad, etc.
  - Se levantan y configuran los servicios necesarios después de tener base solida.
  - Se aplica DataOps: Garantiza que los pipelines de datos y el resultado en capa consumo es el correcto.
  - Si el proyecto es IA, se aplicarán técnicas MLOps, LLMOps: Se generan derivas y se automatiza el proceso de reentrenamiento y despliegue. Ejemplo: Canary .

Fase III (Opcional)

- El equipo se encarga de analizar los pipelines, métricas, derivas para garantizar que todo funciona correctamente.
- Supervisan que el proceso de reentrenamiento y despliegue se ha realizado con éxito. En el caso de falla, se corregirá la problemática.

Yield Predictivo

Resumen del Caso de Uso 8: Yield Predictivo
Este caso de uso se centra en evolucionar la estrategia de monetización de SSMAS, pasando de una optimización reactiva a una **optimización predictiva** para maximizar los ingresos de cada impresión publicitaria. El objetivo es dejar de basarse únicamente en las condiciones actuales del mercado y, en su lugar, anticipar el valor de una impresión antes de que entre en la subasta.

¿Cómo funciona?
La solución consiste en desarrollar un modelo de machine learning avanzado que predice el Coste Por Mil (CPM) esperado para cada oportunidad de anuncio individual.

1.Preparación de Datos con AWS Glue:

- Se utilizan los enormes volúmenes de datos históricos de SSMAS (más de 7.000 millones de impresiones mensuales) como base para el entrenamiento.
- Un trabajo de **AWS Glue** procesa y enriquece estos datos, realizando la **ingeniería de características (feature engineering)**. Esto implica extraer variables relevantes como el tipo de contenido, la fuente del tráfico, la hora del día, la geografía del usuario y las señales de demanda del mercado.

2.Modelo Predictivo con Amazon SageMaker (XGBoost):

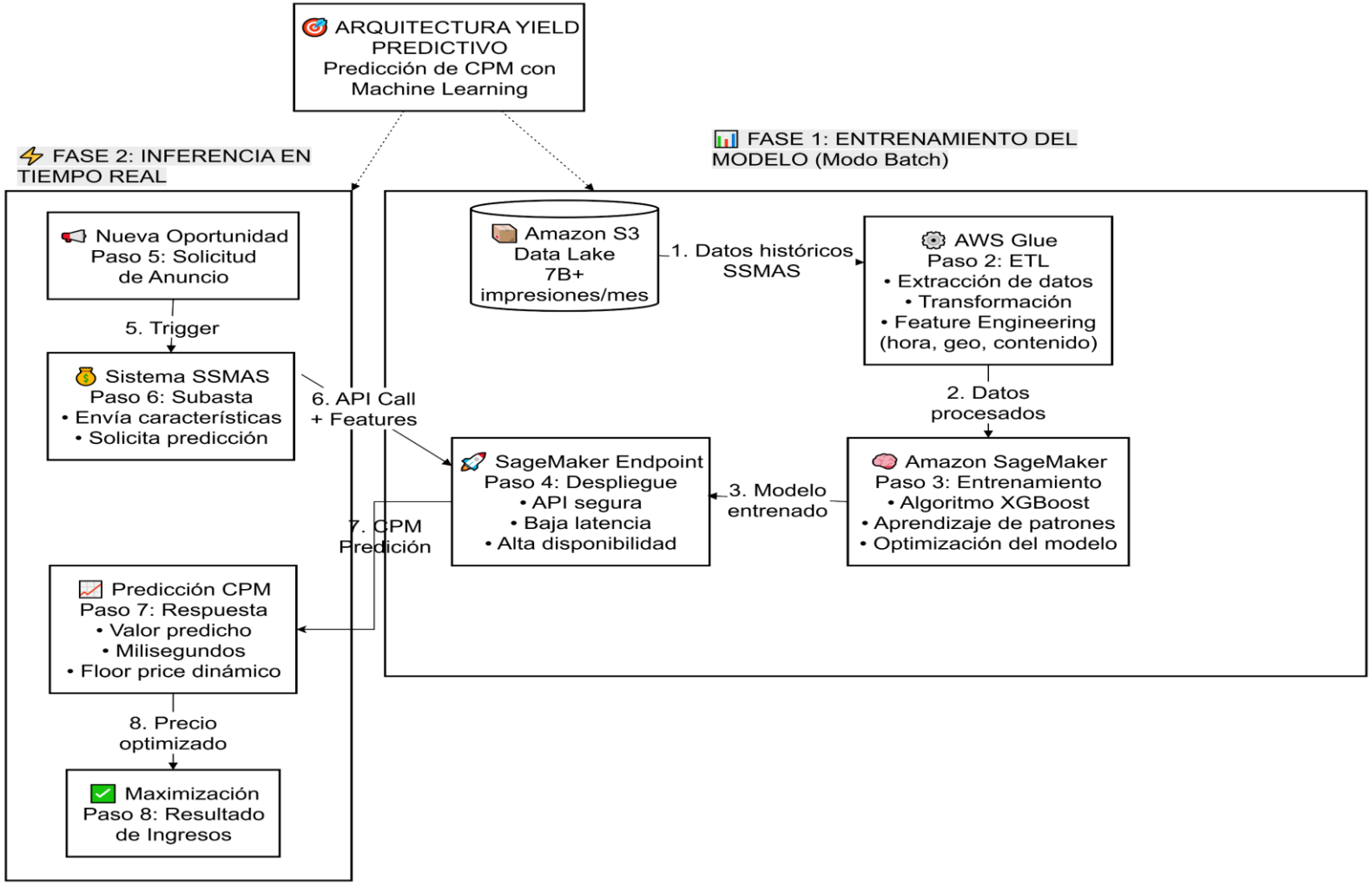
- Se entrena un modelo de regresión utilizando el algoritmo **XGBoost (eXtreme Gradient Boosting)** en Amazon SageMaker. XGBoost es una implementación de árboles de decisión potenciados por gradiente, muy eficaz y popular para trabajar con datos tabulares y realizar predicciones de alta precisión.
- El modelo aprende de los patrones históricos para predecir el CPM más probable que una impresión específica puede alcanzar en la subasta, basándose en sus características.

3.Aplicación de Precios Dinámicos:

- La predicción del modelo se utiliza para establecer **precios mínimos dinámicos (dynamic floor pricing)** para cada subasta en tiempo real.
- Si el modelo predice un CPM alto para una impresión, se establece un precio mínimo más elevado para asegurar que no se venda por menos de su valor. Si predice un CPM bajo, el precio mínimo se puede ajustar a la baja para aumentar la probabilidad de que se venda (mejorar la tasa de relleno o *fill rate*).

Beneficio Empresarial Clave
El principal beneficio es la **maximización de los ingresos totales (yield)** mediante la fijación de precios inteligentes y dinámicos. En lugar de aplicar precios mínimos estáticos, este enfoque permite a SSMAS tomar una decisión de precios óptima para cada una de los miles de millones de impresiones que gestiona, equilibrando perfectamente el precio de venta con la probabilidad de que el anuncio se muestre. Esto refuerza directamente la propuesta de valor central de SSMAS y le proporciona una ventaja tecnológica significativa

Yield Predictivo



Visión Ingram

Yield Predictivo – Fase I – PoC y Desarrollo del MVP

•**Objetivo Principal:** Validar la viabilidad técnica del modelo predictivo en un entorno controlado y construir un MVP funcional que demuestre el valor de negocio.
•**Estimación de Duración:** 4-6 semanas.
•**Actividades Clave:**
- **Configuración del Entorno Aislado:** Despliegue de la infraestructura necesaria en una cuenta de AWS designada (VPC, roles IAM, buckets S3) para garantizar la seguridad y el aislamiento.
- **Ingesta y Procesamiento de Datos:** Carga de la muestra de datos históricos en un bucket de S3. Desarrollo y ejecución de un job de AWS Glue para realizar el ETL (Extracción, Transformación y Carga) y la ingeniería de características (feature engineering).
- **Entrenamiento y Evaluación del Modelo:** Utilización de Amazon SageMaker para entrenar un modelo de regresión con el algoritmo XGBoost. Se evaluará su rendimiento utilizando métricas estándar (ej. RMSE, MAE) y se realizarán ajustes de hiperparámetros si es necesario.
- **Análisis de Resultados y Documentación:** Análisis del poder predictivo del modelo y su potencial impacto en el negocio. Documentación completa del proceso, el código y los resultados obtenidos.

•**Entregables:**
- Modelo de Machine Learning entrenado y guardado como artefacto.
- Informe detallado del rendimiento del modelo y sus métricas.
- Código fuente del job de AWS Glue y de los notebooks de SageMaker, versionado en un repositorio.
- Presentación de resultados y validación del MVP

Yield Predictivo – Fase I – PoC y Desarrollo del MVP

Objetivo Principal: Validar la viabilidad técnica del modelo predictivo en un entorno controlado y construir un MVP funcional que demuestre el valor de negocio.

Plan de Trabajo

Tarea	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
Fase I - Yield Predictivo													
Kick Off													
Configuración Entorno Aislado													
Ingesta y Procesamiento Datos													
Entrenamiento y evaluación del modelo													
Analisis de Resultados y Documentación													
Revisión y aprobación del cliente													

Hitos y actividades

- **Configuración del Entorno Aislado:** Despliegue de la infraestructura necesaria en una cuenta de AWS designada (VPC, roles IAM, buckets S3) para garantizar la seguridad y el aislamiento.
- **Ingesta y Procesamiento de Datos:** Carga de la muestra de datos históricos en un bucket de S3. Desarrollo y ejecución de un job de AWS Glue para realizar el ETL (Extracción, Transformación y Carga) y la ingeniería de características (feature engineering).
- **Entrenamiento y Evaluación del Modelo:** Utilización de Amazon SageMaker para entrenar un modelo de regresión con el algoritmo XGBoost. Se evaluará su rendimiento utilizando métricas estándar (ej. RMSE, MAE) y se realizarán ajustes de hiperparámetros si es necesario.
- **Análisis de Resultados y Documentación:** Análisis del poder predictivo del modelo y su potencial impacto en el negocio. Documentación completa del proceso, el código y los resultados obtenidos.

Entregables

	• Modelo de Machine Learning entrenado y guardado como artefacto.
	• Informe detallado del rendimiento del modelo y sus métricas.
	• Código fuente del job de AWS Glue y de los notebooks de SageMaker, versionado en un repositorio.
	• Presentación de resultados y validación del MVP

Visión Ingram

Yield Predictivo – Fase II – Operativización y Aplicación MLOps

Objetivo Principal: Integrar el MVP en un entorno productivo, seguro y escalable, automatizando el ciclo de vida del modelo para garantizar su mantenimiento y rendimiento a largo plazo.
Estimación de Duración: 6-8 semanas.
Actividades Clave:

- **Diseño de Arquitectura Productiva:** Definición de la arquitectura final siguiendo las mejores prácticas de AWS (ej. Landing Zone, redes, seguridad avanzada).
- **Despliegue del Modelo como Endpoint:** Creación de un endpoint de inferencia en tiempo real en Amazon SageMaker, asegurando baja latencia y alta disponibilidad para responder a las solicitudes del sistema SSMAS.
- **Creación de Pipelines de MLOps:** Implementación de pipelines automatizados con AWS Step Functions o SageMaker Pipelines para orquestar el re-entrenamiento, evaluación y despliegue del modelo.
- **Integración con Sistema SSMAS:** Desarrollo de la llamada API desde el sistema de subastas de SSMAS hacia el endpoint de SageMaker para obtener las predicciones de CPM en tiempo real.
- **Monitorización:** Configuración de dashboards en Amazon CloudWatch para monitorizar el rendimiento del endpoint (latencia, errores) y la calidad de las predicciones.

Entregables:

- Endpoint de inferencia en tiempo real, seguro y escalable.
- Pipeline de re-entrenamiento y despliegue completamente automatizado (CI/CD para ML).
- Dashboard de monitorización de métricas operativas y de negocio.
- Integración funcional con el sistema de subastas.

Visión Ingram

Yield Predictivo – Fase II – Operativización y Aplicación MLOps

Objetivo Principal: Integrar el MVP en un entorno productivo, seguro y escalable, automatizando el ciclo de vida del modelo para garantizar su mantenimiento y rendimiento a largo plazo.

Plan de Trabajo		
Hitos y actividades		• **Diseño de Arquitectura Productiva:** Definición de la arquitectura final siguiendo las mejores prácticas de AWS (ej. Landing Zone, redes, seguridad avanzada). • **Despliegue del Modelo como Endpoint:** Creación de un endpoint de inferencia en tiempo real en Amazon SageMaker, asegurando baja latencia y alta disponibilidad para responder a las solicitudes del sistema SSMAS. • **Creación de Pipelines de MLOps:** Implementación de pipelines automatizados con AWS Step Functions o SageMaker Pipelines para orquestar el re-entrenamiento, evaluación y despliegue del modelo. • **Integración con Sistema SSMAS:** Desarrollo de la llamada API desde el sistema de subastas de SSMAS hacia el endpoint de SageMaker para obtener las predicciones de CPM en tiempo real. • **Monitorización:** Configuración de dashboards en Amazon CloudWatch para monitorizar el rendimiento del endpoint (latencia, errores) y la calidad de las predicciones
Entregables		• Endpoint de inferencia en tiempo real, seguro y escalable.
		• Pipeline de re-entrenamiento y despliegue completamente automatizado (CI/CD para ML).
		• Dashboard de monitorización de métricas operativas y de negocio.
		• Integración funcional con el sistema de subastas.

Visión Ingram

Yield Predictivo – Fase III – Optimización y Mantenimiento Continuo (Opcional)

Objetivo Principal: Garantizar que el modelo siga siendo preciso y relevante a lo largo del tiempo, y optimizar continuamente los recursos y costes de la solución.
Estado: Proceso continuo.
Actividades Clave:

- **Monitorización Activa de Deriva (Model Drift):** Supervisión constante de la desviación del modelo para detectar cuándo las predicciones empiezan a perder precisión debido a cambios en los patrones de datos.
- **Ciclos de Re-entrenamiento y Despliegue:** Ejecución automática o manual de los pipelines de MLOps para re-entrenar el modelo con datos nuevos y desplegar la nueva versión sin tiempo de inactividad (ej. despliegues Canary).
- **Optimización de Costes:** Revisión periódica de los recursos de AWS utilizados para ajustar tamaños de instancias y optimizar el gasto.
- **A/B Testing de Modelos:** Implementación de pruebas para comparar diferentes versiones del modelo en producción y asegurar que cada nueva versión mejora los resultados.

Entregables:

- Informes periódicos de rendimiento del modelo en producción.
- Nuevas versiones del modelo desplegadas de forma controlada.
- Informes de optimización de costes.
- .

Visión Ingram

Yield Predictivo – Fase III – Optimización y Mantenimiento Continuo (Opcional)

Objetivo Principal: Garantizar que el modelo siga siendo preciso y relevante a lo largo del tiempo, y optimizar continuamente los recursos y costes de la solución.

Hitos y actividades	**Monitorización Activa de Deriva (Model Drift): Supervisión constante de la desviación del modelo para detectar cuándo las predicciones empiezan a perder precisión debido a cambios en los patrones de datos.** **Ciclos de Re-entrenamiento y Despliegue: Ejecución automática o manual de los pipelines de MLOps para re-entrenar el modelo con datos nuevos y desplegar la nueva versión sin tiempo de inactividad (ej. despliegues Canary).** **Optimización de Costes: Revisión periódica de los recursos de AWS utilizados para ajustar tamaños de instancias y optimizar el gasto.** **A/B Testing de Modelos: Implementación de pruebas para comparar diferentes versiones del modelo en producción y asegurar que cada nueva versión mejora los resultados.** **esultados.**	
Entregables		Informes periódicos de rendimiento del modelo en producción.
		Nuevas versiones del modelo desplegadas de forma controlada.
		Informes de optimización de costes.

Agente de Adops Autónomo

Resumen del Caso de Uso 10: Agente de AdOps Autónomo
Este caso de uso es el más avanzado y busca transformar radicalmente las operaciones de SSMAS, pasando de un modelo reactivo a un sistema proactivo de **auto-reparación (self-healing)**.
El objetivo es automatizar por completo el ciclo de monitorización, diagnóstico y resolución de problemas operativos 24/7.
¿Cómo funciona?
La solución consiste en construir un "ingeniero de AdOps digital" que puede actuar de forma autónoma cuando se detecta una anomalía en la plataforma.
1.Detección Automática (Amazon CloudWatch):

1. El sistema monitoriza continuamente métricas de negocio críticas (como la tasa de relleno, el RPM de un editor, la latencia de un socio, etc.).
2. Cuando una métrica cruza un umbral predefinido, una **Alarma de CloudWatch** se dispara automáticamente.

2.Orquestación Inteligente (Amazon Bedrock Agents):

1. La alarma activa un **Agente de Bedrock**, que actúa como el "cerebro" del sistema. Este agente está programado con las instrucciones y los manuales de procedimiento (*runbooks*) que seguiría un ingeniero humano.

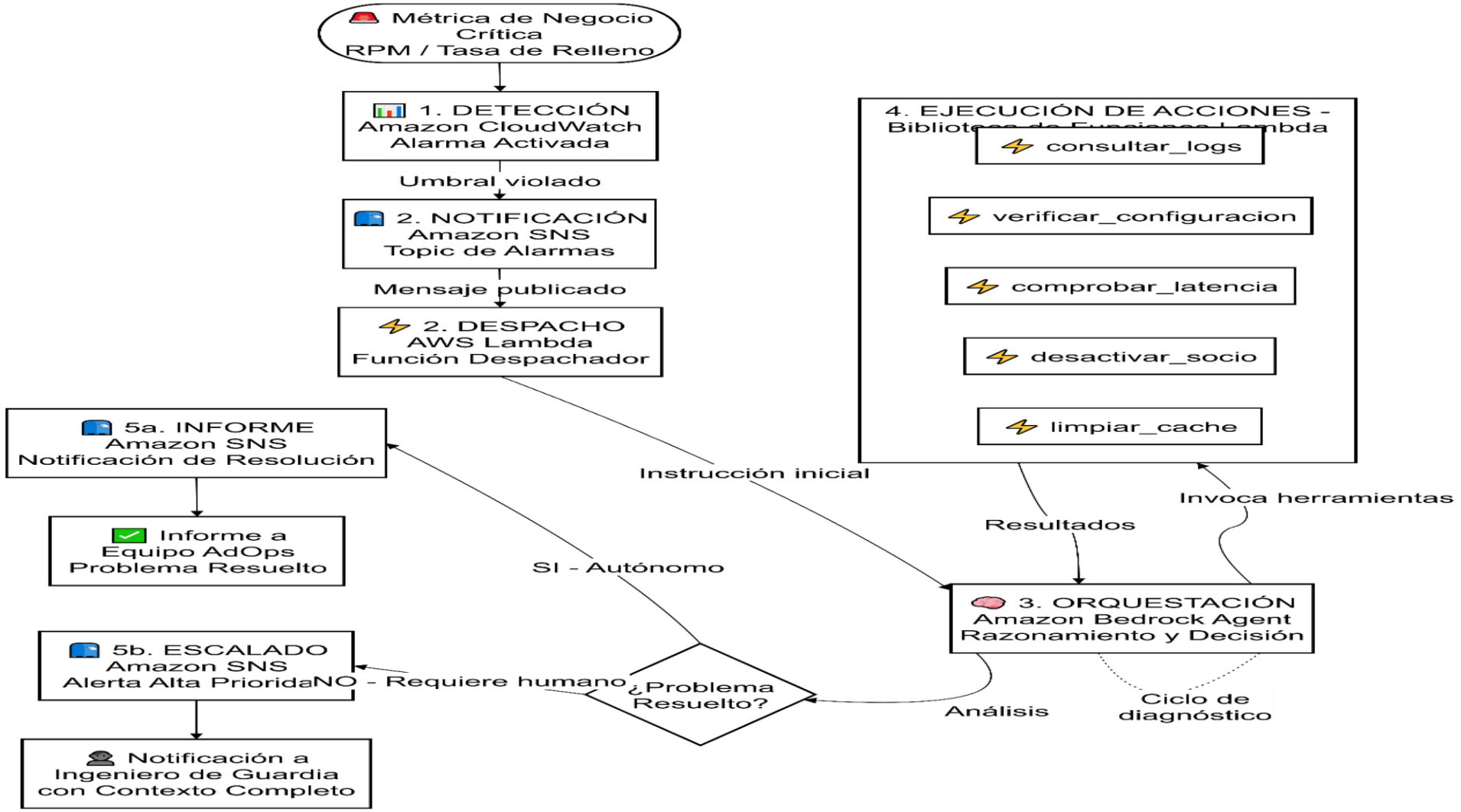
3.Diagnóstico y Remediación (AWS Lambda):

1. El agente comienza a ejecutar una secuencia de diagnóstico utilizando un conjunto de "herramientas" que son, en realidad, funciones de **AWS Lambda**.
2. Por ejemplo, si hay una caída de ingresos, el agente podría invocar una función Lambda para `verificar_configuracion_editor`, luego otra para `analizar_logs_de_errores` y otra para `comprobar_salud_demand_partners`.
3. Una vez que el agente identifica la causa raíz (por ejemplo, un socio de demanda está respondiendo con lentitud), invoca otra función Lambda de remediación para, por ejemplo, `desactivar_temporalmente_socio`.

4.Notificación y Escalado (Amazon SNS):

1. Si el agente resuelve el problema, envía un informe detallado de la resolución al equipo de AdOps a través de **Amazon SNS**.
2. Si el agente no puede resolver el incidente, utiliza SNS para escalar el problema a un ingeniero humano de guardia, proporcionando todo el contexto de la investigación que ya ha realizado.

Agente de Adops Autónomo



Agente de AdOps Autónomo – Fase I – PoC y Desarrollo del MVP

Objetivo Principal: Construir y validar un agente autónomo básico capaz de detectar, diagnosticar y resolver **un tipo de anomalía específica** en un entorno controlado, demostrando la viabilidad de la solución.

Plan de Trabajo		
Hitos y actividades	• Definición del Escenario MVP: Selección de una métrica de negocio crítica (ej. caída del RPM de un editor específico) y definición del manual de procedimiento (runbook) que el agente deberá seguir. • Configuración de la Detección: Creación de una alarma en Amazon CloudWatch que se dispare cuando la métrica seleccionada cruce el umbral definido. • Desarrollo de Herramientas (AWS Lambda): Implementación de un conjunto básico de funciones Lambda que servirán como las "herramientas" del agente para diagnóstico (ej. verificar_configuracion_editor, comprobar_latencia_socio) y remediación (ej. desactivar_socio_temporalmente). • Creación del Agente (Amazon Bedrock): Configuración de un Agente de Bedrock, definiendo sus instrucciones, el acceso a las herramientas Lambda y la lógica de razonamiento inicial. • Pruebas End-to-End: Integración del flujo completo (Alarma -> Agente -> Lambdas -> Notificación) y realización de pruebas simulando la anomalía para validar el comportamiento autónomo.	
Entregables		**Un agente de Bedrock funcional para el escenario definido.**
		Biblioteca de funciones Lambda (herramientas) versionada en un repositorio.
		Informe de resultados de las pruebas de simulación.
		Presentación y demo del MVP.

Agente de AdOps Autónomo– Fase II – Operativización y Aplicación MLOps

Objetivo Principal: Expandir el conocimiento y las habilidades del agente para manejar múltiples tipos de anomalías y refinar su lógica de decisión para un entorno productivo.

Plan de Trabajo		
Hitos y actividades	• Ampliación de la Biblioteca de Herramientas: Desarrollo de nuevas funciones Lambda para cubrir más escenarios de diagnóstico y acciones de remediación. • Refinamiento del Agente: Mejora de las instrucciones y la lógica de razonamiento del Agente de Bedrock para permitirle manejar escenarios más complejos y tomar decisiones más matizadas. • Implementación de Pipelines de CI/CD: Creación de un proceso automatizado para el despliegue y actualización de las funciones Lambda (las "herramientas"). • Monitorización y Trazabilidad: Configuración de un sistema de logging avanzado (ej. Amazon CloudWatch Logs) para tener una traza completa de las decisiones y acciones del agente. • Integración Controlada en Producción: Despliegue del agente en modo "sombra" (shadow mode) o con aprobaciones manuales antes de permitir la remediación totalmente autónoma.	
Entregables		**Agente con capacidad para gestionar múltiples anomalías.**
		Pipeline de CI/CD para las herramientas del agente.
		Dashboard de monitorización de la actividad y eficacia del agente.
		Plan de despliegue progresivo en el entorno productivo.

Yield Predictivo – Fase III – Optimización y Mantenimiento Continuo (Opcional)

Objetivo Principal: Asegurar que el agente mantenga su eficacia a lo largo del tiempo, se adapte a los cambios en la plataforma y opere de forma rentable

Hitos y actividades	• **Revisión de Rendimiento: Análisis periódico de la tasa de éxito del agente en la resolución de problemas y de los casos que requieren escalado a humanos.** • **Actualización de Runbooks: Modificación de las instrucciones y herramientas del agente para reflejar nuevos procedimientos operativos o cambios en la arquitectura.** • **Optimización de Costes: Monitorización del coste de ejecución de las Lambdas y las llamadas al Agente de Bedrock para asegurar un uso eficiente.** • **Entrenamiento Continuo: Utilización de los casos escalados a humanos como fuente de conocimiento para enseñar al agente a manejar nuevas situaciones en el futuro.**	
Entregables		**Informes de rendimiento y eficacia del agente.**
		Actualizaciones periódicas del agente con nuevas capacidades.
		Informes de optimización de costes.

Agente de Adops Autónomo