AWS re:Invent

XALDIGITAL

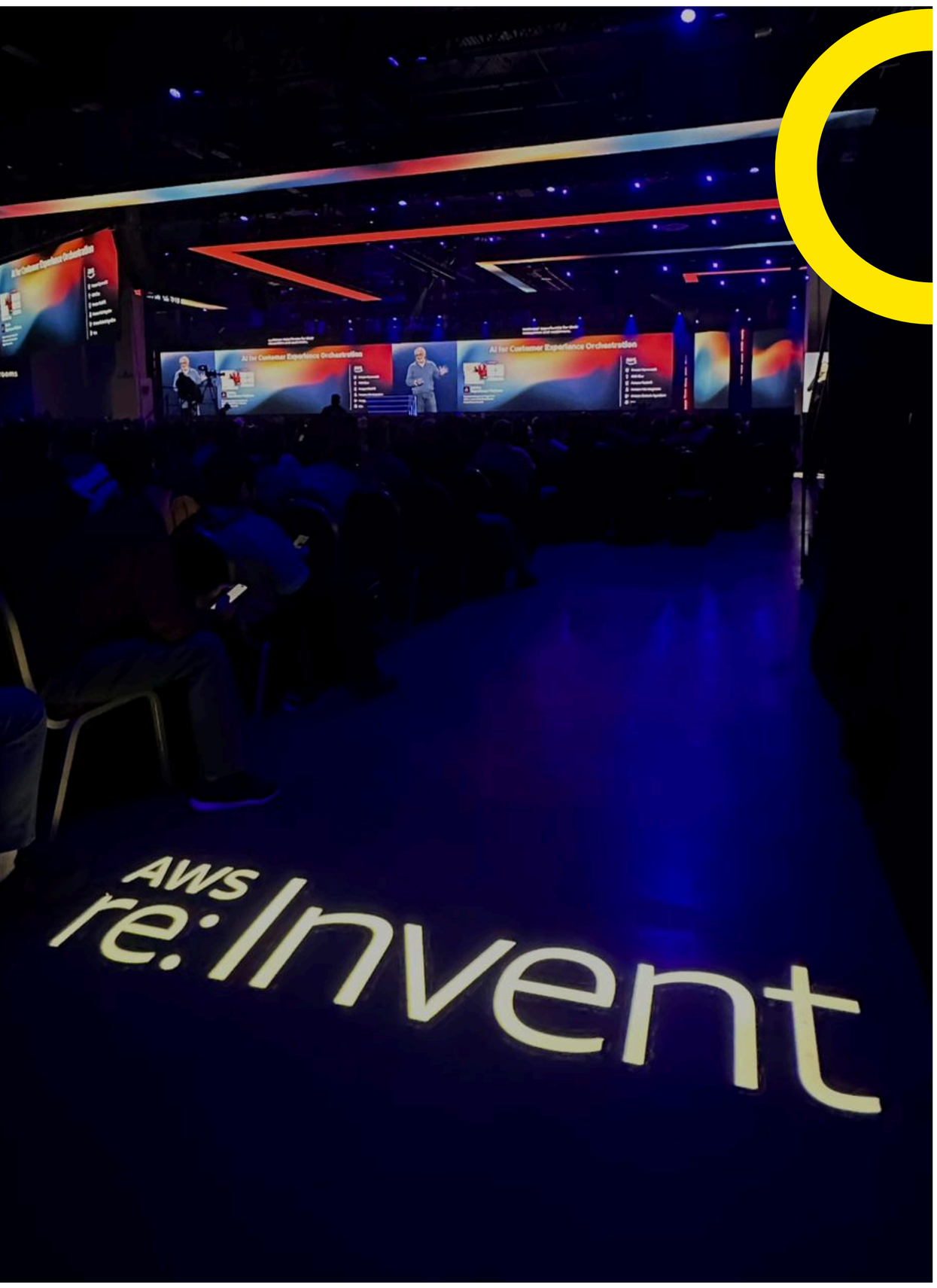
aws PARTNER Premier Tier Services

DOSSIER TÉCNICO/COMERCIAL UNIFICADO

AWS RE:INVENT 2025

Versión completa con 100% de contenidos integrados

¿QUÉ CAMBIA PARA XALDIGITAL?

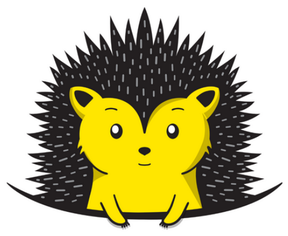


AWS anunció una combinación revolucionaria de infraestructura de IA, nuevos modelos generativos, agentes gobernables, datos listos para IA, y modernización acelerada con impacto directo en nuestro portafolio de soluciones y capacidades técnicas. Esta evolución representa una oportunidad única para posicionar a **XalDigital como líder** en transformación digital impulsada por inteligencia artificial en el mercado latinoamericano.

El ecosistema AWS ha evolucionado significativamente, integrando componentes que antes requerían múltiples herramientas y proveedores. Esta consolidación no solo simplifica la arquitectura técnica, sino que también reduce la complejidad operativa y los costos asociados a la gestión de infraestructura distribuida.

Para **XalDigital**, esto significa capacidad de entregar soluciones end-to-end más competitivas y con tiempos de implementación reducidos.

PILARES TECNOLÓGICOS



INFRAESTRUCTURA DE IA ACELERADA

Trainium3, nuevas GPUs NVIDIA y mejoras sustanciales en inferencia. AWS AI Factories llevan clusters de IA dedicados directamente a datacenters de clientes. Project Rainier introduce una nueva arquitectura base para ejecutar modelos multimodales a gran escala, optimizando eficiencia energética, latencia y rendimiento global.

DATOS LISTOS PARA IA

Amazon S3 Vectors proporciona soporte nativo para almacenar vectores con hasta 90% menos costo. S3 Tables introduce tablas transaccionales nativas en S3 integradas con Athena, Redshift y EMR. Mejoras críticas incluyen S3 Batch más rápido, replicación automática multi-región y FSx NetApp optimizado.

MODELOS NOVA Y PERSONALIZACIÓN PROFUNDA

Amazon Nova y Nova 2 representan modelos multimodales frontier. Nova 2 Lite está optimizado para chatbots y copilotos. Nova 2 Pro maneja razonamiento avanzado, análisis de video y migraciones de software. Nova Forge permite crear modelos propios basados en checkpoints oficiales.

AGENTES CONFIABLES Y GOBERNADOS

Amazon Bedrock AgentCore introduce políticas determinísticas, evaluaciones de calidad continuas, memoria episódica y observabilidad nativa. El foco está en seguridad, gobernanza empresarial y cumplimiento normativo estricto para sectores regulados.

MODERNIZACIÓN ACELERADA CON IA

AWS Transform Custom acelera la modernización 5× más rápido que sistemas legacy tradicionales. Agentes especializados en Security, DevOps, FinOps y MLOps. Reducción de hasta 70% en costos de mantenimiento y licencias de software propietario.

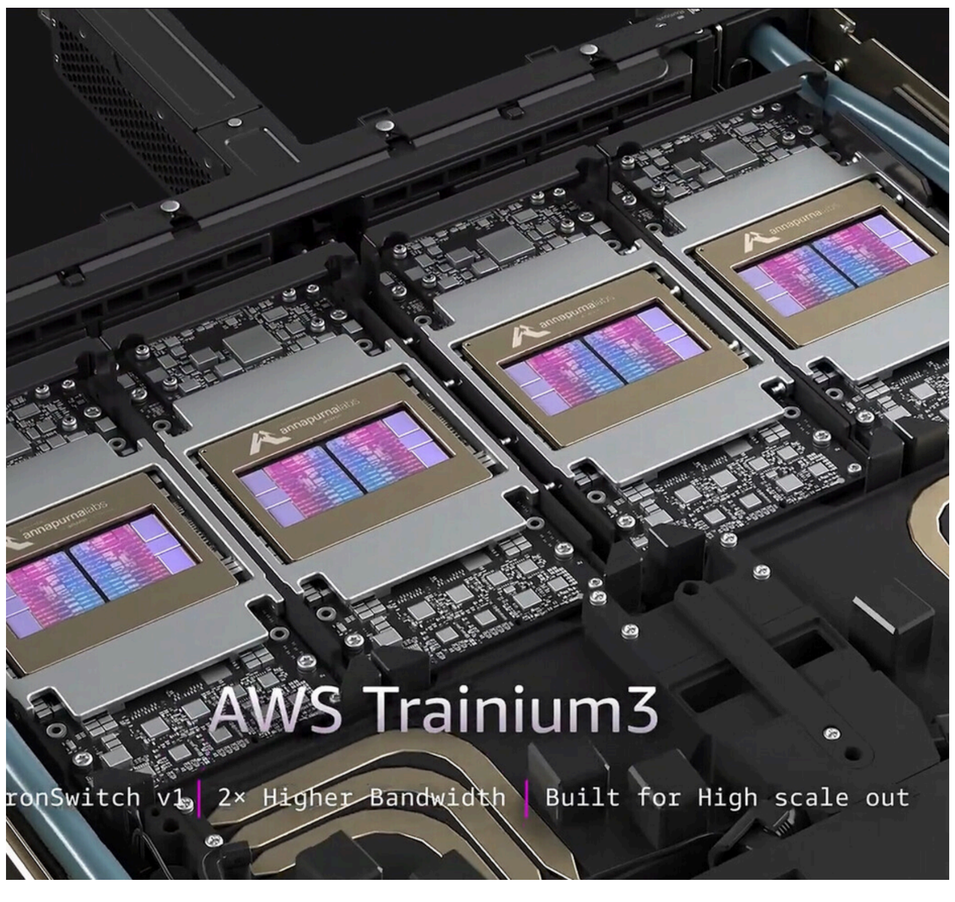
La integración de estos cinco pilares tecnológicos representa un cambio fundamental en cómo las empresas pueden abordar la transformación digital. Para XalDigital, el desafío y la oportunidad radican en traducir estas capacidades técnicas en propuestas de valor concretas que resuelvan problemas de negocio específicos de nuestros clientes en sectores como banca, retail, manufactura y gobierno.

INFRAESTRUCTURA DE IA Y CÓMPUTO

La evolución de la infraestructura de cómputo para inteligencia artificial marca un antes y un después en las capacidades de entrenamiento e inferencia de modelos. Trainium3 representa la tercera generación de chips diseñados específicamente por AWS para cargas de trabajo de machine learning, ofreciendo mejoras sustanciales en rendimiento y eficiencia energética que impactan directamente en el TCO (Total Cost of Ownership) de proyectos de IA empresarial.

TRAINIUM3: NUEVA GENERACIÓN

- Incremento de +4× en rendimiento comparado con Trainium2
- Mejor escalabilidad para entrenamiento de modelos grandes (>100B parámetros)
- Interconexión optimizada para clusters distribuidos
- Soporte nativo para formatos de precisión mixta



PROJECT RAINIER: ARQUITECTURA FUNDAMENTAL

Project Rainier constituye el nuevo stack de infraestructura diseñado para ejecutar modelos de IA a escala masiva. Representa la base tecnológica sobre la cual se construye la familia Nova y optimiza aspectos críticos como latencia en inferencia, eficiencia energética, throughput para cargas multimodales y despliegue distribuido en geografías múltiples.

VENTAJA COMPETITIVA

La combinación de Trainium3 y Project Rainier reduce costos de entrenamiento hasta 60% vs. alternativas en la nube.

INSTANCIAS EC2 DE ÚLTIMA GENERACIÓN

C8A (COMPUTE OPTIMIZED)

Optimizadas para cargas computacionales intensivas: modelado financiero, simulaciones, análisis de riesgo, procesamiento de transacciones de alta frecuencia y renderizado.

X8 / X8AEZ (MEMORY INTENSIVE)

Diseñadas para bases de datos en memoria, SAP HANA, analytics en tiempo real, cache distribuido y aplicaciones que requieren grandes volúmenes de RAM.

M8AZN (GENERAL PURPOSE)

Balance óptimo entre cómputo, memoria y red. Ideales para aplicaciones empresariales, microservicios, entornos de desarrollo y cargas de trabajo mixtas.

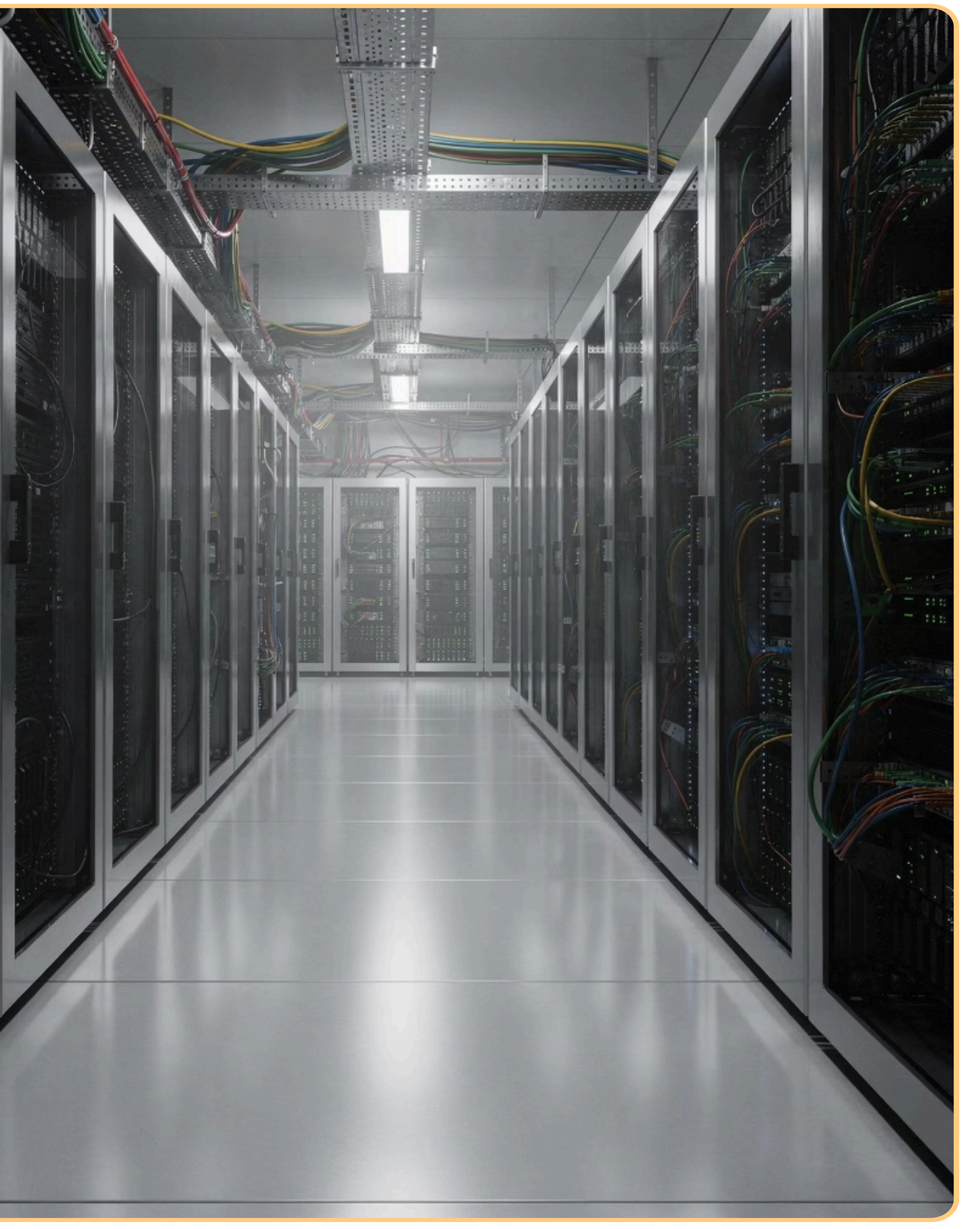
IMPACTO ESTRATÉGICO PARA XALDIGITAL

La disponibilidad de esta nueva generación de infraestructura nos permite desarrollar soluciones diferenciadas para nuestros clientes. Podemos ofrecer **"AI Training Clusters as a Service",** proporcionando clusters de IA dedicados con gestión completa, desde el aprovisionamiento hasta la operación continua y optimización de costos. Esta capacidad es especialmente relevante para clientes en sectores financiero, retail y manufactura que están desarrollando modelos propietarios de machine learning.

La integración de estas nuevas instancias en nuestras arquitecturas de landing zones permite acelerar la implementación de soluciones de pricing dinámico, análisis de riesgo crediticio, revenue management, optimización de cadenas de suministro y analítica predictiva intensiva. Los beneficios en términos de latencia, throughput y costo por inferencia son sustanciales, especialmente en escenarios que requieren procesamiento en tiempo real de grandes volúmenes de datos.

AWS AI FACTORIES

AWS AI Factories representa un cambio paradigmático en cómo las organizaciones pueden desplegar infraestructura de inteligencia artificial a escala empresarial. Este modelo lleva la infraestructura completa de IA —incluyendo red de alta velocidad, almacenamiento distribuido y cómputo especializado— directamente a los datacenters de los clientes, eliminando restricciones de latencia y cumplimiento normativo que históricamente han limitado la adopción de soluciones cloud en sectores altamente regulados.



SECTORES OBJETIVO

Las AI Factories son especialmente relevantes para organizaciones en banca, gobierno, telecomunicaciones y salud que enfrentan restricciones regulatorias estrictas sobre dónde pueden residir y procesarse los datos sensibles.

CARACTERÍSTICAS PRINCIPALES

- Infraestructura de IA completa instalada on-premises
- Conectividad de baja latencia (sub-milisegundo)
- Alta disponibilidad con redundancia geográfica
- Cumplimiento normativo garantizado
- Gestión híbrida AWS-cliente
- Escalabilidad modular según demanda

ARQUITECTURA Y CAPACIDADES TÉCNICAS

La arquitectura de AI Factories combina hardware especializado de AWS con software de orquestación que permite entrenar e inferir modelos frontier en entornos controlados. La solución incluye clusters de Trainium para entrenamiento, instancias Inferentia para inferencia optimizada, almacenamiento S3 compatible instalado localmente, y conectividad Direct Connect de alta velocidad hacia servicios AWS en la nube pública para escenarios híbridos.

ASSESSMENT Y DISEÑO

Análisis de capacidad y diseño arquitectónico óptimo.

APROVISIONAMIENTO

Instalación y conexión integral del hardware AWS.

IMPLEMENTACIÓN DE WORKLOADS

Despliegue y optimización de modelos de IA.

OPERACIÓN CONTINUA

Monitoreo, seguridad y soporte especializado 24/7.

OPORTUNIDAD COMERCIAL PARA XALDIGITAL

XalDigital abre tres líneas de negocio con el modelo de AI Factories: (1) un Readiness Assessment de 4–6 semanas para evaluar la madurez del cliente y definir el roadmap; el diseño e implementación de AI Factories híbridas en 3–6 meses, integrando infraestructura, conectividad y MLOps; y la operación continua con servicios gestionados de MLOps, seguridad y optimización, generando ingresos recurrentes y una relación estratégica de largo plazo.

LAMBDA MANAGED INSTANCES

Lambda Managed Instances representa una evolución significativa del modelo serverless tradicional, combinando la simplicidad operativa de AWS Lambda con el control y la potencia de cómputo de instancias EC2. Esta innovación elimina la dicotomía histórica entre serverless y compute tradicional, permitiendo a los desarrolladores mantener el modelo de programación event-driven de Lambda mientras acceden a recursos de cómputo más robustos cuando la carga de trabajo lo requiere.

IMPACTO ESTRATÉGICO PARA XALDIGITAL

MODELO HÍBRIDO SERVERLESS-COMPUTE

Las funciones Lambda pueden ejecutarse en instancias EC2 administradas completamente por AWS, sin necesidad de gestión de infraestructura por parte del usuario. El escalado, parcheo, monitoreo y recuperación ante fallos son completamente automáticos.

AUTOSCALING INTELIGENTE

El sistema escala automáticamente según la carga de trabajo, aprovisionando y des-aprovisionando instancias EC2 de manera dinámica. Los algoritmos de scaling predicen patrones de para minimizar cold starts y optimizar costodemandas.

GESTIÓN AUTOMATIZADA COMPLETA

AWS maneja actualizaciones de seguridad, parches del sistema operativo, configuración de red, balanceo de carga y recuperación automática ante fallos. Los desarrolladores solo despliegan código.

CASOS DE USO EMPRESARIALES

Lambda Managed Instances brilla en escenarios que históricamente han sido difíciles para serverless puro. El procesamiento intensivo de datos es uno de los casos más relevantes: transformaciones ETL complejas, procesamiento de imágenes y video, análisis de logs masivos, y generación de reportes que requieren alto CPU o memoria ahora pueden ejecutarse en el modelo serverless sin las limitaciones de timeout o memoria de Lambda estándar.

Los workloads con requisitos de performance altos también se benefician significativamente. Aplicaciones de analytics en tiempo real, procesamiento de streaming de eventos, cálculos financieros complejos y simulaciones Monte Carlo pueden ejecutarse con la latencia y throughput de EC2 pero con la simplicidad operativa de Lambda.

Los pipelines de datos y ETL son otra área de alto impacto. Nuestras soluciones de ingesta, transformación y carga de datos pueden ahora manejar volúmenes significativamente mayores y transformaciones más complejas sin la necesidad de provisionar clusters EMR o Glue pesados. Esto resulta en pipelines más ágiles, económicos y fáciles de mantener, especialmente relevante para clientes en retail y financiero con requisitos intensivos de procesamiento de datos.

DATOS Y ALMACENAMIENTO LISTOS PARA IA

AMAZON S3 VECTORS

Amazon S3 Vectors introduce soporte nativo para almacenamiento y consulta de embeddings vectoriales directamente en S3, eliminando la necesidad de bases de datos vectoriales especializadas para muchos casos de uso. Esta capacidad representa un cambio fundamental en la economía y arquitectura de soluciones de IA generativa, especialmente aquellas que implementan patrones RAG (Retrieval Augmented Generation) a escala empresarial.

ARQUITECTURA Y CAPACIDADES

Reducción del 90% en costos de almacenamiento vectorial
Eliminación de infraestructura de bases de datos vectoriales
Escalabilidad ilimitada sin provisioning de capacidad
Integración nativa con SageMaker y Bedrock
Durabilidad 99.999999999% de S3
•ersionado y auditoría incorporados

ROI INMEDIATO

Clientes con >100M vectores pueden ahorrar **$50K-$200K** anuales migrando a S3 Vectors.

CASOS DE USO EMPRESARIALES PRIORITARIOS

RAG (RETRIEVAL AUGMENTED GENERATION)

Construcción de sistemas de pregunta-respuesta sobre documentos corporativos. Los embeddings de documentos, contratos, manuales y políticas se almacenan en S3 Vectors y se consultan en tiempo real para enriquecer respuestas de LLMs con contexto específico de la organización.

BÚSQUEDA SEMÁNTICA EMPRESARIAL

Motores de búsqueda que entienden intención y contexto, no solo palabras clave. Aplicable a búsqueda de productos en e-commerce, documentación técnica, knowledge bases y repositorios de código.

AGENTES CON MEMORIA VECTORIAL

Agentes conversacionales que recuerdan interacciones históricas y contexto de usuario. La memoria episódica se almacena como vectores permitiendo recuperación eficiente de conversaciones y contextos relevantes.

INDEXACIÓN MULTIMODAL

Indexación y búsqueda de documentos, imágenes, audio y video usando embeddings multimodales. Permite búsqueda cross-modal como "encontrar documentos que discutan conceptos similares a esta imagen".

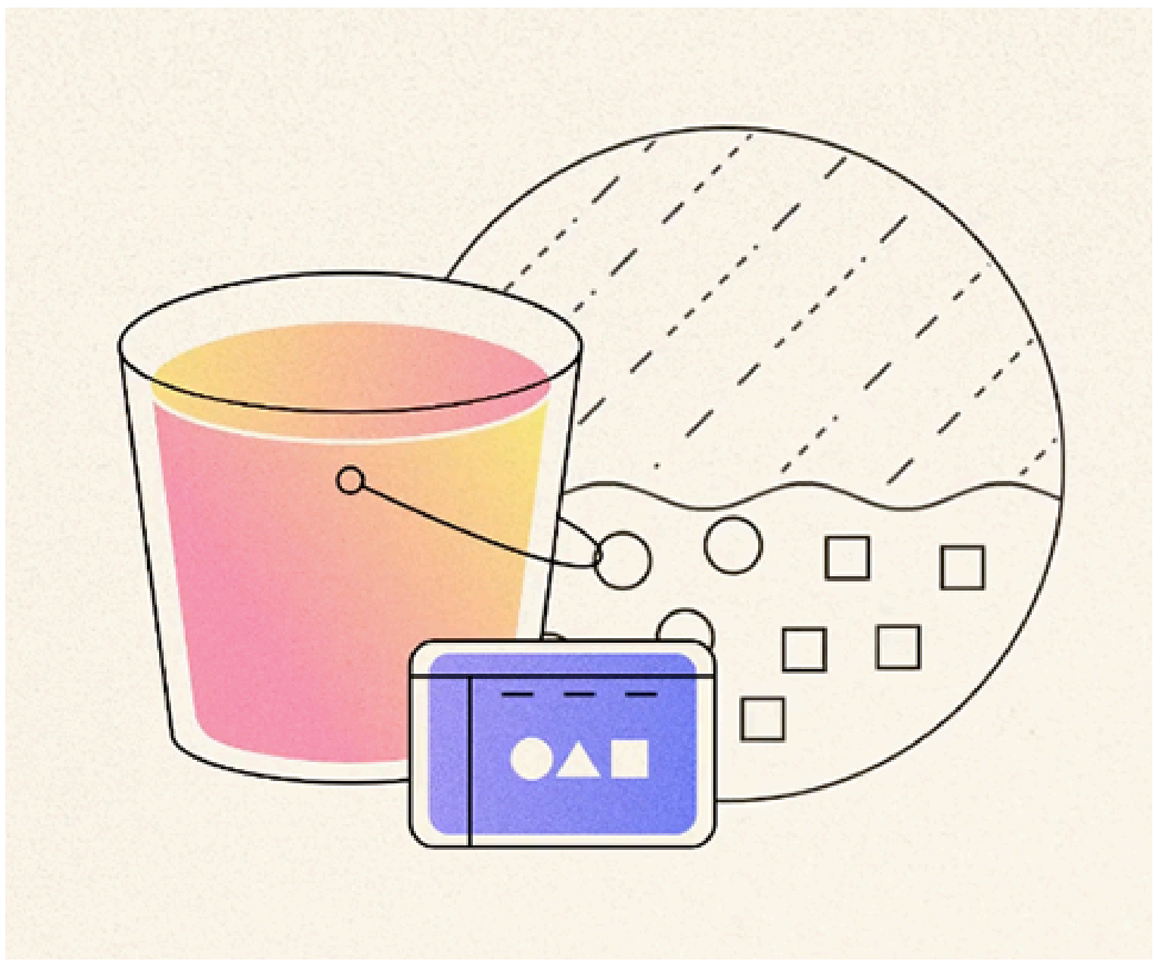
PROPUESTA DE VALOR XALDIGITAL

Podemos desarrollar un acelerador empaquetado: "RAG sobre S3 para Enterprise". Este acelerador incluye arquitectura de referencia, pipeline de ingesta de documentos, generación de embeddings usando modelos optimizados (Amazon Titan o Cohere), implementación de S3 Vectors con indexación automática, y API de consulta integrada con Bedrock. El tiempo de implementación típico es 3-4 semanas con casos de uso específicos del cliente.

La conversión de data lakes tradicionales a AI-ready lakes es otra oportunidad significativa. Muchos clientes tienen data lakes en S3 con terabytes o petabytes de datos no estructurados (logs, documentos, imágenes) que no están siendo aprovechados para IA. Podemos ofrecer servicios de enriquecimiento donde procesamos estos datos históricos, generamos embeddings vectoriales, los almacenamos en S3 Vectors y construimos capacidades de búsqueda semántica y RAG sobre ellos. Este servicio se monetiza por volumen de datos procesados más una componente recurrente de mantenimiento y actualización continua.

S3 TABLES

S3 Tables introduce un modelo de almacenamiento tabular transaccional directamente en Amazon S3, eliminando la necesidad de clusters de procesamiento dedicados para manipulación de datos estructurados. Esta capacidad representa una convergencia entre data lakes y data warehouses, proporcionando las ventajas de escalabilidad y costo de S3 con las garantías ACID y performance de consulta de sistemas tabulares tradicionales.



ARQUITECTURA Y DIFERENCIACIÓN TÉCNICA

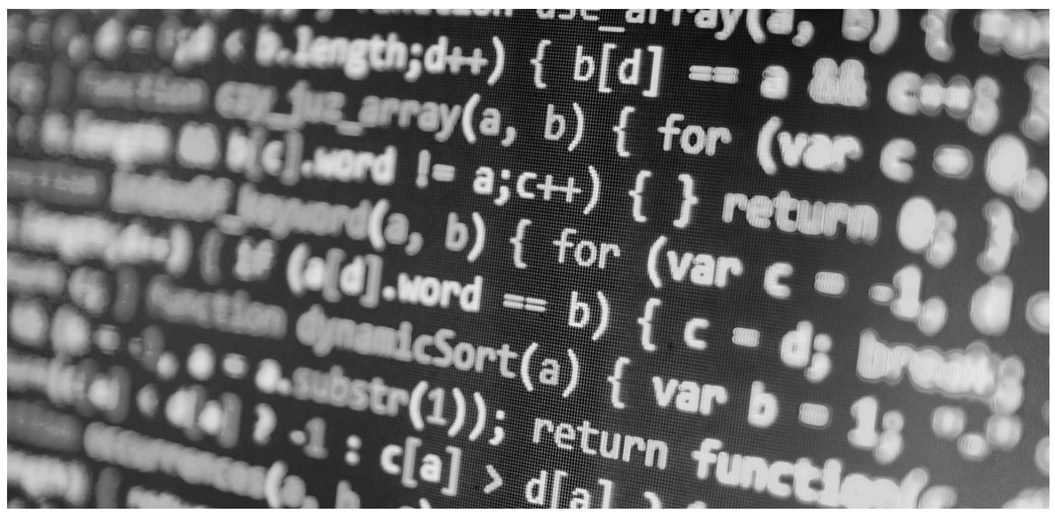
A diferencia de formatos como Parquet o ORC que son solo almacenamiento columnar, S3 Tables incorpora un motor de transacciones que maneja versionado, concurrencia y consistencia. Cada tabla mantiene un registro de transacciones que permite rollback, time travel queries, y auditoría completa de cambios. El esquema es enforced a nivel de almacenamiento, previniendo errores de schema drift comunes en data lakes tradicionales.

ARQUITECTURA Y DIFERENCIACIÓN TÉCNICA

- **Amazon Athena:** consultas SQL interactivas sin provisioning
- **Amazon EMR:** procesamiento distribuido Spark/Hive
- **Amazon Redshift**: consultas federadas de alta performance
- **AWS Glue**: ETL gestionado y catalogación automática
- **SageMaker**: lectura directa para feature engineering
- •**Herramientas BI**: integración con Tableau, Power BI, Looker

ARQUITECTURA Y DIFERENCIACIÓN TÉCNICA

- Versionado automático de datos con time travel
- Rollback a puntos específicos en el tiempo
- Concurrencia optimista para escrituras
- Schema evolution sin downtime
- Auditoría completa de modificaciones
- Particionamiento automático inteligente



CASOS DE USO TRANSFORMACIONALES

GOBIERNO DE DATOS ROBUSTO

Organizaciones en sectores regulados (banca, salud, gobierno) necesitan auditoría completa de acceso y modificación de datos. S3 Tables proporciona registro inmutable de cambios, control de versiones y capacidad de demostrar compliance con regulaciones como GDPR, SOX o Basel III.

MACHINE LEARNING REPRODUCIBLE

El entrenamiento de modelos de ML requiere datasets inmutables y reproducibles. S3 Tables permite a los científicos de datos versionar datasets, referenciar versiones específicas en experimentos y reproducir resultados meses o años después con los datos exactos usados originalmente.

ANALYTICS HISTÓRICO EFICIENTE

Time travel queries permiten analizar cómo han evolucionado los datos sin necesidad de mantener copias completas. Los analistas pueden ejecutar consultas como "estado de inventario al cierre de cada mes en el último año" de manera eficiente.

IMPACTO EN OFERTAS DE XALDIGITAL

S3 Tables nos permite simplificar significativamente nuestras arquitecturas de Data Governance. Muchos proyectos actuales requieren infraestructura compleja con Delta Lake, Apache Iceberg o Hudi para lograr transaccionalidad sobre S3. Con S3 Tables, esta complejidad desaparece, reduciendo el tiempo de implementación de 8-12 semanas a 3-4 semanas y eliminando la necesidad de gestionar metastores externos y motores de procesamiento dedicados.

Nuestras soluciones de Data Lake Modernization se benefician dramáticamente. Podemos ofrecer un camino de **migración más simple y económico** para clientes con data lakes legacy basados en Hive o repositorios relacionales que quieren modernizar. El proceso típico incluye assessment de datos existentes, diseño de esquemas optimizados para S3 Tables, migración con versionado completo y habilitación de capacidades de time travel y auditoría. El ROI es claro: mismos o mejores tiempos de consulta, reducción del 50-70% en costos de infraestructura de procesamiento, y mejora sustancial en gobierno de datos.

La **reducción de complejidad en pipelines ETL** es otro beneficio estratégico. Muchos pipelines actuales requieren múltiples pasos de staging, validación y movimiento de datos entre sistemas para asegurar consistencia. Con S3 Tables, los pipelines se simplifican porque las garantías transaccionales están incorporadas en el almacenamiento mismo. Esto resulta en pipelines más confiables, fáciles de mantener y económicos de operar.

IA GENERATIVA: AMAZON NOVA & NOVA FORGE

Amazon Nova representa la entrada de AWS en el espacio de modelos de lenguaje frontier desarrollados internamente, competiendo directamente con GPT-4, Claude y Gemini. La familia Nova no solo incluye los modelos base sino también Nova Forge, una plataforma de entrenamiento y customización que permite a las organizaciones crear modelos propietarios manteniendo control completo sobre datos, arquitectura y despliegue.

AMAZON NOVA (GENERACIÓN 1)

Modelo multimodal frontier capaz de procesar y generar texto, imágenes y código. Optimizado para razonamiento complejo, análisis contextual profundo y generación de contenido creativo. Comparable en capacidades a GPT-4 con ventajas en integración nativa con servicios AWS y costos optimizados.

NOVA 2 LITE

Modelo ligero optimizado para latencia sub-segundo y alto throughput. Diseñado específicamente para chatbots de servicio al cliente, asistentes virtuales, copilotos de código y aplicaciones que requieren respuestas rápidas con consumo mínimo de recursos. Costo por token 70% menor que Nova estándar.

NOVA 2 PRO

Modelo avanzado con capacidades extendidas de razonamiento, análisis de video frame-by-frame, comprensión de diagramas técnicos complejos y asistencia en migraciones de software legacy. Incluye fine-tuning específico para tareas empresariales como análisis de contratos, revisión de código y documentación técnica automática.

CAPACIDADES MULTIMODALES AVANZADAS

Nova 2 Pro destaca particularmente en análisis de video, una capacidad diferenciada vs. competidores. Puede procesar streams de video completos, identificar eventos específicos, generar resúmenes temporales, transcribir y traducir audio, y extraer insights de contenido visual complejo. Esto abre casos de uso en seguridad (análisis de video vigilancia), retail (comportamiento en tienda), manufactura (control de calidad visual) y media (catalogación automática de contenido).

CAPACIDADES MULTIMODALES AVANZADAS

Nova Forge representa un cambio paradigmático en cómo las empresas pueden desarrollar modelos de IA propietarios. Tradicionalmente, entrenar un modelo frontier requería equipos especializados, infraestructura masiva (miles de GPUs durante meses) y presupuestos de millones de dólares. Nova Forge democratiza este proceso proporcionando checkpoints pre-entrenados de los modelos Nova como punto de partida, reduciendo el tiempo y costo de entrenamiento en un orden de magnitud.

CHECKPOINT NOVA

Modelo base pre-entrenado con capacidades generales. Punto de partida con billones de tokens procesados y conocimiento broad.

FINE-TUNING GESTIONADO

Nova Forge ejecuta entrenamiento optimizado en infraestructura AWS con MLOps automatizado, monitoreo continuo y validación de calidad.

DATOS PROPIETARIOS

El cliente proporciona datasets específicos de su industria, procesos internos, documentación y casos de uso únicos.

MODELO CUSTOMIZADO

Modelo propietario del cliente, optimizado para sus casos de uso, deployable en entornos gobernados con SLAs empresariales.

GOBERNANZA Y SEGURIDAD EMPRESARIAL

Nova Forge incluye capas robustas de seguridad y gobernanza críticas para clientes enterprise. Los datos de entrenamiento nunca salen del entorno del cliente (VPC aislado o AI Factory). El modelo resultante es propiedad exclusiva del cliente con políticas de acceso granulares. Las capacidades de MLOps incluyen versionado automático de modelos, A/B testing gestionado, monitoreo de drift y reentrenamiento programado. La observabilidad nativa proporciona métricas de performance, latencia, costo por inferencia y detección de anomalías.

ESTRATEGIA COMERCIAL PARA XALDIGITAL

Nova Forge nos permite lanzar **"Modelos XalDigital by Industry",** una línea de modelos verticalizados desarrollados en colaboración con clientes anchor en cada sector.

- **Financiero**: Detección de fraude, scoring crediticio, análisis de riesgo, cumplimiento automático (AML/KYC) y reportes normativos. Ventaja clave: conocimiento de regulaciones LATAM (CNBV México, CMF Chile, SFC Colombia).
- **Manufactura/Movilidad:** Mantenimiento predictivo, optimización de supply chain, planificación de producción y logística last-mile mediante procesamiento de datos IoT, telemetría vehicular y variables externas.
- **Monetización:** Setup fee inicial (implementación, ingesta de datos, fine-tuning) + consumo mensual recurrente por volumen de inferencias. Genera revenue predecible y escalable alineado al valor del cliente.
- **Diferenciación:** Únicos con modelos optimizados para contexto, idioma, regulaciones y casos de uso específicos de Latinoamérica.

AGENTES, GOBERNANZA Y MODERNIZACIÓN

AMAZON BEDROCK AGENTCORE

Amazon Bedrock AgentCore representa la evolución de agentes conversacionales de chatbots reactivos a agentes autónomos capaces de ejecutar tareas complejas de manera confiable, segura y auditable. La plataforma incorpora cuatro pilares fundamentales que abordan los desafíos críticos de desplegar agentes de IA en entornos empresariales: políticas determinísticas, memoria contextual, evaluaciones continuas de calidad y observabilidad profunda.

POLICY IN AGENTCORE

Sistema de políticas que define reglas determinísticas sobre las acciones permitidas del agente. Las políticas se expresan en lenguaje declarativo y se evalúan antes de cada acción. Incluye controles de acceso a datos, límites de costos, restricciones de tiempo de respuesta y reglas de escalamiento.

AGENTCORE EVALUATIONS

Framework de evaluación continua que mide calidad de respuestas, adherencia a políticas, latencia, costo y satisfacción del usuario. Incluye dashboards con métricas de negocio y técnicas, alertas automáticas y capacidades de A/B testing.

AGENTCORE MEMORY

Memoria episódica que permite a los agentes recordar interacciones previas, contexto de usuario y aprendizajes de sesiones anteriores. Almacenada como vectores en S3 para recuperación eficiente. Incluye políticas de retención, anonimización y eliminación conforme a GDPR.

OBSERVABILIDAD NATIVA

Integración con CloudWatch y X-Ray proporcionando traces distribuidos, logs estructurados, métricas de performance en tiempo real y debugging de conversaciones problemáticas. Esencial para operar agentes en producción con SLAs exigentes.

ENFOQUE EN SECTORES REGULADOS

AgentCore está diseñado específicamente para sectores con requisitos estrictos de seguridad, gobernanza y cumplimiento. En banca, los agentes pueden manejar consultas de balance, movimientos, pagos y transferencias mientras aseguran que cada acción cumple políticas de AML, limites transaccionales y consentimiento explícito del usuario. Los logs detallados y la trazabilidad completa facilitan auditorías regulatorias.

En telecomunicaciones, agentes de soporte pueden diagnosticar problemas técnicos, ejecutar pruebas de red, aprovisionar servicios y gestionar reclamos manteniendo trazabilidad completa de cada acción y decisión. Las políticas previenen acciones que podrían afectar servicios críticos sin aprobación humana.

Para gobierno, donde la transparencia y accountability son críticas, AgentCore proporciona registro inmutable de decisiones, explicabilidad de razonamiento y capacidad de demostrar que cada respuesta del agente cumplió políticas y regulaciones aplicables.

PROPUESTA DE VALOR XALDIGITAL

Podemos desarrollar una línea de "Agentes Empresariales Gobernados" específicos por industria. Estos agentes pre-configurados incluirían políticas baseline, evaluaciones estándar de calidad, memoria optimizada para casos de uso verticales y dashboards ejecutivos customizados. El tiempo de implementación se reduciría de 12-16 semanas (desarrollo desde cero) a 4-6 semanas (configuración y customización de agentes pre-construidos).

MODERNIZACIÓN CON AWS TRANSFORM CUSTOM

AWS Transform Custom es un acelerador de modernización impulsado por IA que reduce drámáticamente el tiempo, costo y riesgo de migrar aplicaciones legacy a arquitecturas cloud-native. Combina análisis automatizado de código, agentes especializados en diferentes dominios técnicos y frameworks de refactorización guiados por ML, logrando modernizaciones 5× más rápido que enfoques manuales tradicionales.

CAPACIDADES TÉCNICAS FUNDAMENTALES

Transform Custom analiza aplicaciones existentes en múltiples dimensiones: dependencias de código, patrones arquitecturales, integraciones con sistemas externos, esquemas de bases de datos, configuraciones de infraestructura y requisitos de cumplimiento. Este análisis genera un grafo de dependencias completo que informa las estrategias de modernización recomendadas.
Los agentes especializados ejecutan refactorizaciones complejas de manera autónoma. El agente de Security identifica vulnerabilidades, configuraciones inseguras y falta de encriptación, generando código remediado automáticamente. El agente de DevOps refactoriza pipelines de CI/CD, genera infraestructura como código (Terraform/CloudFormation) y establece estrategias de deployment blue-green o canary.

ESPECIALIZACIÓN POR DOMINIO TÉCNICO

AGENTE SECURITY

Identifica vulnerabilidades (OWASP Top 10), configuraciones inseguras, secretos hardcodeados y dependencias obsoletas con CVEs conocidos. Genera código remediado, implementa secrets management con AWS Secrets Manager y configura security groups mínimo-privilegio.

AGENTE DEVOPS

Moderniza pipelines CI/CD legacy a CodePipeline/CodeBuild, genera infraestructura como código, implementa estrategias de deployment avanzadas (blue-green, canary) y establece observabilidad con CloudWatch/X-Ray. Automatiza rollback y testing de integración.

AGENTE FINOPS

Analiza costos actuales de infraestructura on-prem y proyecta TCO en AWS. Identifica oportunidades de ahorro (rightsizing, Reserved Instances, Savings Plans, S3 Intelligent-Tiering) y genera reportes ejecutivos de business case con alarmas de presupuesto y dashboards de cost allocation.

AGENTE MLOPS

Moderniza pipelines de ML/AI legacy, migra modelos a SageMaker con A/B testing y monitoreo de drift, implementa feature stores y gobernanza, establece reentrenamiento automático y genera dashboards de performance (accuracy, latency, costo por inferencia).

BENEFICIOS ECONÓMICOS CUANTIFICABLES

Reducción de costos de licencias: Aplicaciones legacy con licencias caras de SQL Server, Oracle o WebSphere migran a alternativas open-source (PostgreSQL, MySQL) o servicios AWS (RDS, Aurora), reduciendo costos 60-80%. Cliente con $500K anuales en licencias Oracle puede reducir a $100K en Aurora, ahorrando $400K anuales.
Costos de infraestructura: Servidores on-prem sobredimensionados (15-30% de capacidad utilizada) se reemplazan por instancias EC2 rightsized con autoscaling, reduciendo costos 40-60%. Servicios serverless (Lambda, Fargate) reducen costos hasta 70% vs. servidores dedicados.
Reducción de costos de mantenimiento: Aplicaciones legacy consumen recursos enormes: equipos especializados caros, documentación perdida, testing manual intensivo y riesgo alto de fallos. Post-modernización, el mantenimiento se reduce 50-70%: infraestructura gestionada por AWS, CI/CD automatizado, testing automatizado y arquitecturas modulares fáciles de modificar.

VELOCIDAD DE MIGRACIÓN

Aplicaciones que tomarían 6-8 meses manualmente se modernizan en 6-10 semanas con Transform Custom.

OFERTA XALDIGITAL TRANSFORM LAB

Establecemos un "XalDigital Transform Lab" como unidad especializada que ejecuta modernizaciones usando Transform Custom. Proceso típico de cuatro fases: Discovery (2-3 semanas) donde analizamos la aplicación y generamos assessment con Transform; Design (2-3 semanas) donde definimos arquitectura cloud-native target y estrategia de migración; Transform (4-8 semanas) donde ejecutamos refactorización guiada por agentes y migramos datos; Optimize (2-4 semanas post-launch) donde ejecutamos rightsizing y cost optimization.

Casos de uso prioritarios:
Migración SQL Server a RDS/Aurora/Linux: Particularmente relevante dado el fin de soporte de Windows Server 2012 y SQL Server 2012. Transform Custom automatiza conversión de stored procedures, triggers y queries T-SQL a dialectos PostgreSQL o MySQL.
Transformación de monolitos a microservicios: Transform analiza el monolito, identifica bounded contexts, genera propuesta de descomposición en microservicios y refactoriza código con agentes especializados. Resultado: aplicaciones más ágiles, deployments independientes, mejor escalabilidad y menor riesgo de cambios.

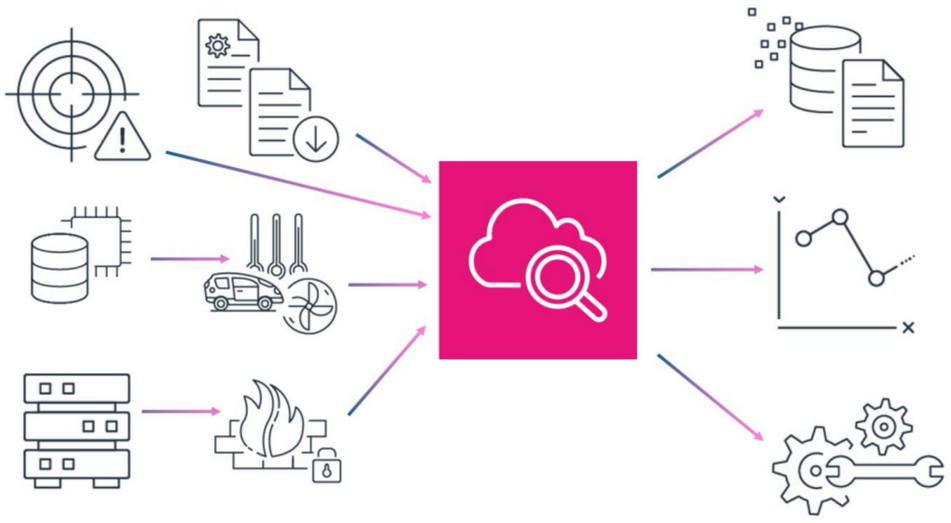
SEGURIDAD, OBSERVABILIDAD Y OPERACIÓN

Los anuncios de re:Invent incluyen mejoras sustanciales en las capacidades de seguridad, observabilidad y operación que son fundamentales para ejecutar cargas de trabajo de IA a escala empresarial. Estas capacidades, aunque menos vistosas que los nuevos modelos de IA, son críticas para asegurar que las soluciones sean seguras, confiables, performantes y económicas en producción.

SECURITY HUB Y GUARDDUTY EXTENDIDOS

AWS Security Hub ha sido reforzado con capacidades de detección de amenazas específicas para cargas de IA/ML. El servicio ahora identifica configuraciones inseguras en notebooks SageMaker, buckets S3 con datasets de entrenamiento públicos, endpoints de modelos sin autenticación y uso de modelos en Bedrock sin políticas de acceso apropiadas. La integración con GuardDuty proporciona detección de comportamiento anómalo: consultas inusuales a modelos de IA, exfiltración potencial de datos de entrenamiento y uso no autorizado de recursos costosos de IA.

Amazon **GuardDuty** extiende su cobertura a contenedores ECS/Fargate y instancias EC2 con detección de runtime mejorada. Identifica actividad de cryptomining, comunicaciones con IPs maliciosas conocidas, escalamiento de privilegios y exfiltración de datos. Para cargas de IA, detecta patrones como entrenamiento no autorizado de modelos, acceso sospechoso a datasets sensibles y uso de recursos de inferencia fuera de horarios normales.



CLOUDWATCH UNIFIED DATA STORE

CloudWatch introduce un Unified Data Store que consolida logs, métricas y traces en un repositorio único con capacidades de consulta unificadas. Anteriormente, estos tres tipos de telemetría requerían APIs diferentes y consultas en silos. El nuevo modelo permite correlación automática: dado un spike de latencia (métrica), el sistema puede mostrar automáticamente los logs y traces relevantes del mismo periodo, acelerando dramáticamente el troubleshooting.

1 DETECCIÓN

Alarma de CloudWatch detecta latencia >2s en endpoint de inferencia. Automáticamente correlaciona con logs y traces del periodo.

2 ANÁLISIS

Dashboard unificado muestra que latencia correlaciona con spike en tamaño de input. Traces revelan bottleneck en preprocesamiento de datos.

3 REMEDIACIÓN

Equipo implementa caching de preprocesamiento. CloudWatch valida que latencia vuelve a <500ms. Incident se cierra.

4 PREVENCIÓN

Sistema aprende el patrón. Alarma predictiva se configura para detectar situaciones similares antes de impacto a usuarios.

IMPLICACIONES OPERACIONALES

Estas capacidades de seguridad y observabilidad mejoradas reducen sustancialmente el Mean Time To Resolution (MTTR) de incidentes. Clientes reportan reducciones de 60-80% en tiempo de troubleshooting gracias a la correlación automática de telemetría. En producción, donde cada minuto de downtime tiene costo directo, esta reducción se traduce en savings cuantificables y mejor experiencia de usuario.

Los costos de herramientas de observabilidad también se optimizan. Muchas organizaciones operan stacks complejos con Splunk, Datadog, New Relic y herramientas adicionales, con costos que pueden alcanzar cientos de miles de dólares anuales. La consolidación en CloudWatch Unified Data Store puede reducir estos costos 50-70% mientras proporciona capacidades equivalentes o superiores.

PROPUESTA XALDIGITAL: SECURITY & OBSERVABILITY FAST TRACK

Incluye: Implementación de Security Hub con políticas por industria, activación de GuardDuty, configuración de CloudWatch Unified Data Store con dashboards ejecutivos/técnicos, alarmas inteligentes con ML Insights y capacitación operativa.

Valor inmediato: Detección proactiva de amenazas, troubleshooting acelerado via telemetría correlacionada, consolidación de herramientas de observabilidad, y base segura para operaciones de IA en producción.

ROADMAP RECOMENDADO PARA XALDIGITAL

La implementación exitosa de estas capacidades requiere un roadmap estructurado que balancee experimentación temprana, desarrollo de capacidades internas y escalamiento de ofertas comerciales. El siguiente roadmap de 12 meses proporciona un camino pragmático para maximizar el valor de los anuncios de re:Invent 2025.

0-90 DÍAS: EXPERIMENTACIÓN Y VALIDACIÓN

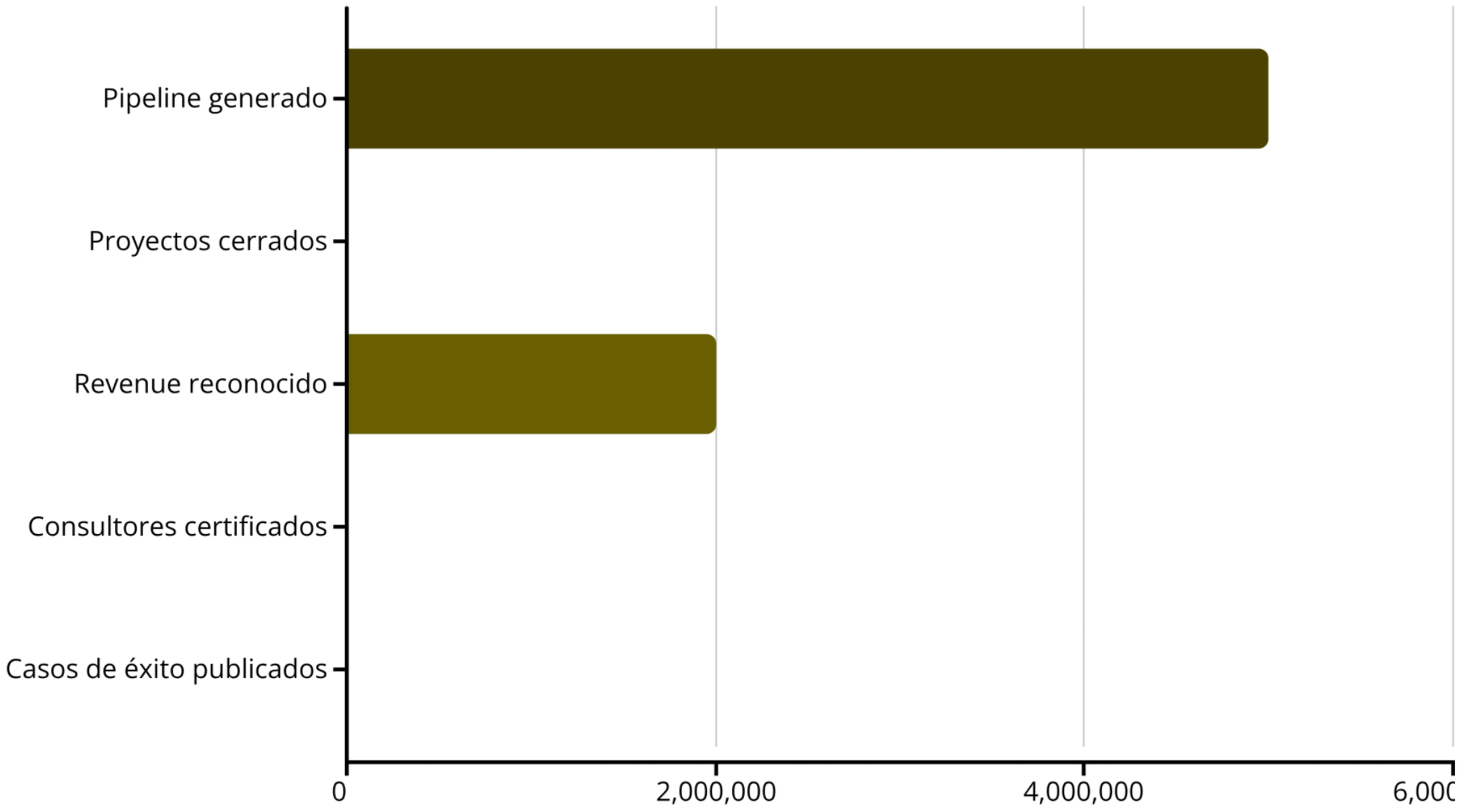
Ejecutar pilotos internos con S3 Vectors, Nova 2 y AgentCore para validar capacidades técnicas y desarrollar conocimiento profundo. Ajustar propuestas comerciales incorporando nuevas capacidades con pricing y casos de uso específicos. Seleccionar 3 design partners estratégicos (uno por sector: banca, retail, manufactura) para pilotos colaborativos con co-inversión y feedback continuo.

90-180 DÍAS: DESARROLLO DE ACELERADORES

Desarrollar AI Factory Assessment como oferta empaquetada con metodología, templates y herramientas. Construir aceleradores 1.0: RAG sobre S3 con arquitectura de referencia y código reutilizable, Modernización con Transform Custom con playbooks por tecnología legacy, y Agents Foundation con AgentCore pre-configurado. Ejecutar habilitación técnica interna intensiva en Nova, AgentCore y Transform Custom con certificaciones AWS y hands-on workshops.

180-365 DÍAS: ESCALAMIENTO Y DIFERENCIACIÓN

Lanzar modelos XalDigital verticalizados usando Nova Forge: retail, financiero y manufactura con datasets iniciales de design partners. Lanzamiento comercial de Agents-as-a-Service con SLAs empresariales, pricing por consumo y soporte 24/7. Establecer integraciones profundas con Project Rainier para clientes estratégicos requiriendo máxima performance y escala. Desarrollar casos de estudio documentados de pilotos exitosos para marketing y ventas.



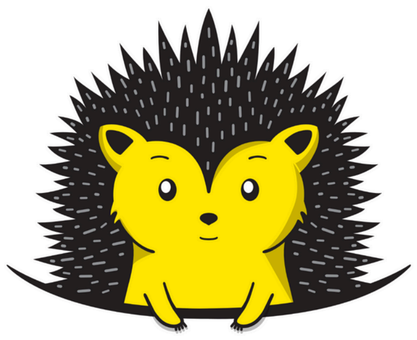
PRÓXIMOS PASOS INMEDIATOS

SEMANA 1-2

- Socializar este dossier con liderazgo
- Identificar design partners candidatos
- Solicitar acceso temprano a nuevos servicios con AWS
- Definir equipo core para pilotos

SEMANA 3-4

- Kick-off de pilotos internos
- Primeras conversaciones con design partners
- Definir estructura de aceleradores
- Planning de habilitación técnica



MES 2-3

- Primeros resultados de pilotos
- Ajuste de propuestas comerciales
- Lanzamiento de habilitación interna
- Firma de design partners

La ejecución disciplinada de este roadmap posicionará a XalDigital como el partner líder en Latinoamérica para transformación digital impulsada por IA generativa, con capacidades diferenciadas, casos de éxito demostrables y relaciones profundas con AWS. El momento de actuar es ahora: la ventana de oportunidad para establecer liderazgo es limitada y la competencia está avanzando rápidamente. Los anuncios de re:Invent 2025 no son solo nuevas features técnicas, son catalizadores para redefinir cómo las empresas en nuestra región construyen, operan y monetizan soluciones de inteligencia artificial.