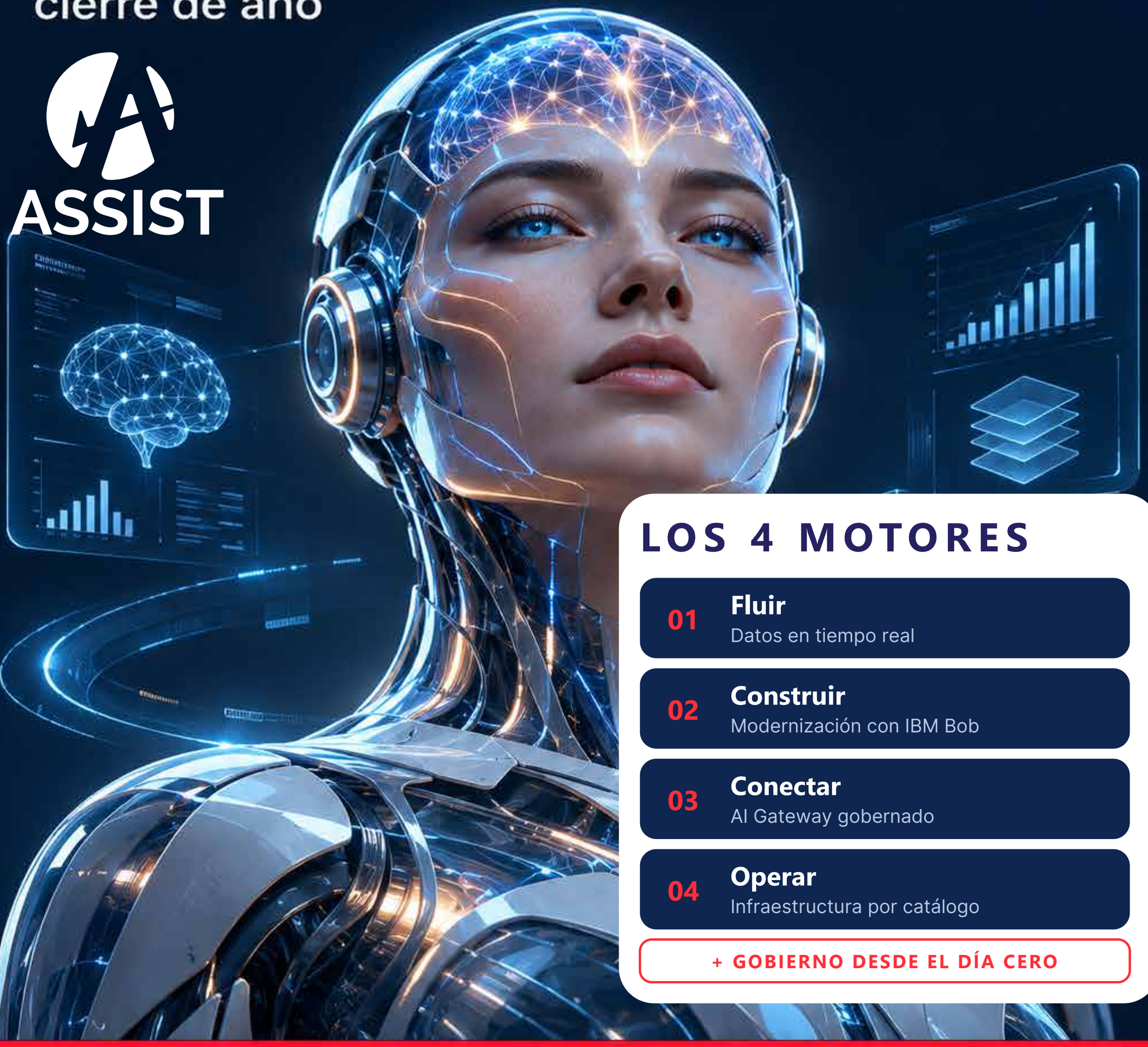
SIN DRAMAS DE DICIEMBRE

Manual para industrializar tu IA antes del cierre de año

ASSIST



LOS 4 MOTORES

01 Fluir
Datos en tiempo real

02 Construir
Modernización con IBM Bob

03 Conectar
AI Gateway gobernado

04 Operar
Infraestructura por catálogo

+ GOBIERNO DESDE EL DÍA CERO

ANTES DE EMPEZAR

La regla no escrita del cierre de año en tecnología

Noviembre y diciembre en tecnología tienen una regla no escrita: nadie quiere tocar producción antes del congelamiento de cambios, todo el mundo anda cuadrando el presupuesto del próximo año y los comités directivos piden ver resultados reales antes de salir a vacaciones.

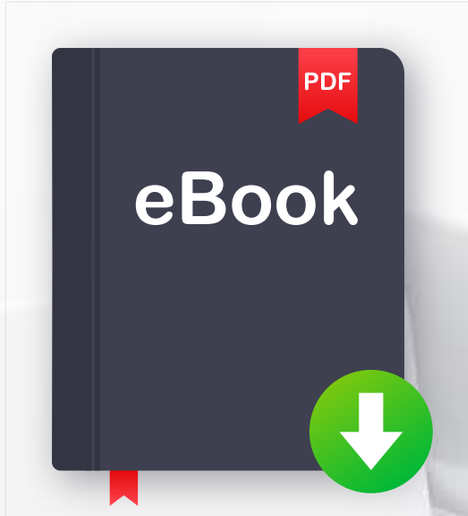
01
Congelamiento
de cambios

02
Presupuesto
del próximo año

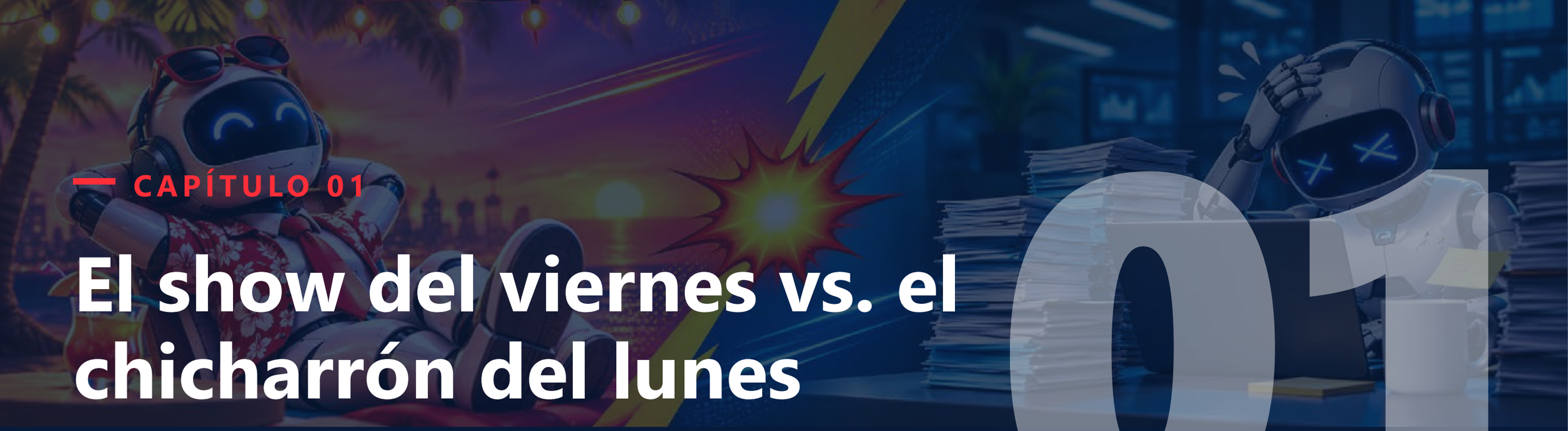
03
Comités
piden resultados

El problema es que muchas empresas llegan al cierre de año con una colección de pilotos de Inteligencia Artificial que se ven divinos en las diapositivas de los viernes, pero que el lunes a las 8:00 a.m. se caen a pedazos apenas reciben carga real.

Este eBook está hecho para que cierres el año con la arquitectura en orden, sin cuentos raros y con la capacidad técnica lista para operar a escala en entornos altamente regulados.







CAPÍTULO 01

El show del viernes vs. el chicharrón del lunes

01

Arranquemos con los números sobre la mesa, sin anestesia.

EL ARRANQUE VS. LA REALIDAD

100% de las organizaciones inicia proyectos de IA → **<5%** logra un éxito escalable

80%

La presión de arriba

de estas iniciativas opera bajo un mandato urgente del CEO.

77%

El descontrol

de los líderes admite que la IA superó por completo sus capacidades de gobernanza corporativa.

40%

La tijera que viene

de los proyectos de IA agéntica serán cancelados antes de 2027, según proyecta Gartner.

11%

Los que sí están listos

de los líderes empresariales está verdaderamente preparado para escalar hoy.

DIAGNÓSTICO

¿Por qué se rompe el sistema cuando pasa a producción?

Cuando una solución falla en diciembre, el deporte corporativo favorito es echarle la culpa al modelo de lenguaje. Pero miremos qué pasa de verdad bajo el capó:

No es falla del modelo

Decir que el algoritmo “alucinó” o que le faltan parámetros es falso a escala empresarial. La inteligencia ya no es el cuello de botella.

No es solo falla de implementación

Pensar que todo se arregla corrigiendo un error puntual del despliegue del viernes es incompleto. Un hito técnico aislado no sostiene la continuidad del negocio.

El verdadero lío es la ejecución

Aquí está el cuello de botella real. El sistema colapsa bajo la carga operativa porque tienes agentes tomando decisiones con datos de hace 24 horas. Además, las APIs están expuestas sin límites de velocidad (rate limits) y nadie tiene idea de cuánto cuesta cada consulta.

Para pasar de pilotos aislados, fragmentación y caos de datos a una operación estandarizada en producción, la ecuación de ingeniería no admite negociaciones:

EJECUCIÓN =

Datos + Integración + Infraestructura + Gobierno

Si uno solo de esos pilares flaquea, todo el sistema se cae.

CAPÍTULO 02

Los 4 motores para que tu arquitectura aguante temporada alta

02

Cuatro motores tecnológicos sostenidos por un eje transversal degobierno.

Si quieres que tu plataforma sobreviva a los picos transaccionales de fin de año y arranque enero sin deudas técnicas, necesitas encender cuatro motores tecnológicos sostenidos por un eje transversal de gobierno.

01 FLUIR

Datos vivos, cero memorias de ayer

Streamhouse · Kafka + Flink · IBM Confluent

02 CONSTRUIR

Modernización con copiloto de verdad

IBM Bob · Plan, Code, Test, Secure

03 CONECTAR

Darle manos a la IA sin descuidar la caja fuerte

AI Gateway · MCP · webMethods / API Connect

04 OPERAR

Infraestructura curada por catálogo, cero inventos

Terraform · Red Hat OpenShift

✳ **EJE TRANSVERSAL: GOBIERNO DESDE EL DÍA CERO**

MOTOR 01 · FLUIR

Datos vivos, cero memorias de ayer

Un agente de IA no puede evaluar si le aprueba un crédito a alguien mirando el saldo que tenía ayer a medianoche. Seguir operando con procesamiento por lotes (batch) es abrirle la puerta de par en par al fraude y a las pérdidas operativas.

EL CAMBIO DE CHIP

Se abandona el esquema de datos muertos por lotes para montar una arquitectura Streamhouse en tiempo real sobre Apache Kafka y Flink, habilitada por IBM Confluent.

Las 5 etapas del flujo continuo

El dato entra y se procesa sin pausas mediante cinco pasos:



LO QUE DICEN LAS CIFRAS · REPORTE DATA STREAMING 2026

72%

de los directivos reconoce que está frenado en IA por tener infraestructuras de datos deficientes.

44% ROI hasta 5x

logra un retorno de inversión gracias al streaming neutral con Kafka y Flink.



MOTOR 02 · CONSTRUIR

Modernización con copiloto de verdad (IBM Bob)

Modernizar un core heredado antes de cerrar el año suena a pesadilla, pero aquí no estamos hablando de un simple autocompletado de texto. IBM Bob entra como un Delivery Partner impulsado por IA que hace enrutamiento inteligente entre modelos:

Frontier **Open Source** **Granite SLM**

El flujo de trabajo

01 Plan → **02 Code** → **03 Test** → **04 Secure**

Código legado (COBOL) → Microservicios con REST API

Control humano obligatorio

La IA no reemplaza al arquitecto humano. Bob analiza dependencias, propone rutas de refactorización y genera las pruebas automatizadas, pero el arquitecto humano es quien revisa y aprueba cada artefacto.

Vigilancia con métricas DORA

Se integra Alerta DORA para monitorear la frecuencia de despliegue, el tiempo de recuperación y la tasa de fallos, asegurando velocidad sin sacrificar estabilidad.



MOTOR 02 · CASOS REALES

Casos reales con números auditados

10x

30 → 3 DÍAS

Caso Blue Pearl

Código heredado

El entendimiento de código JCL/PL/I y el análisis de COBOL para modernizar hacia arquitecturas cloud-native bajó de 30 días a 3 días, logrando una velocidad 10x mayor con 100% de precisión.

-69%

TIEMPO DE EJECUCIÓN

Caso Maximo

Refactorización

La reestructuración estructural redujo el tiempo total de ejecución en un -69%.

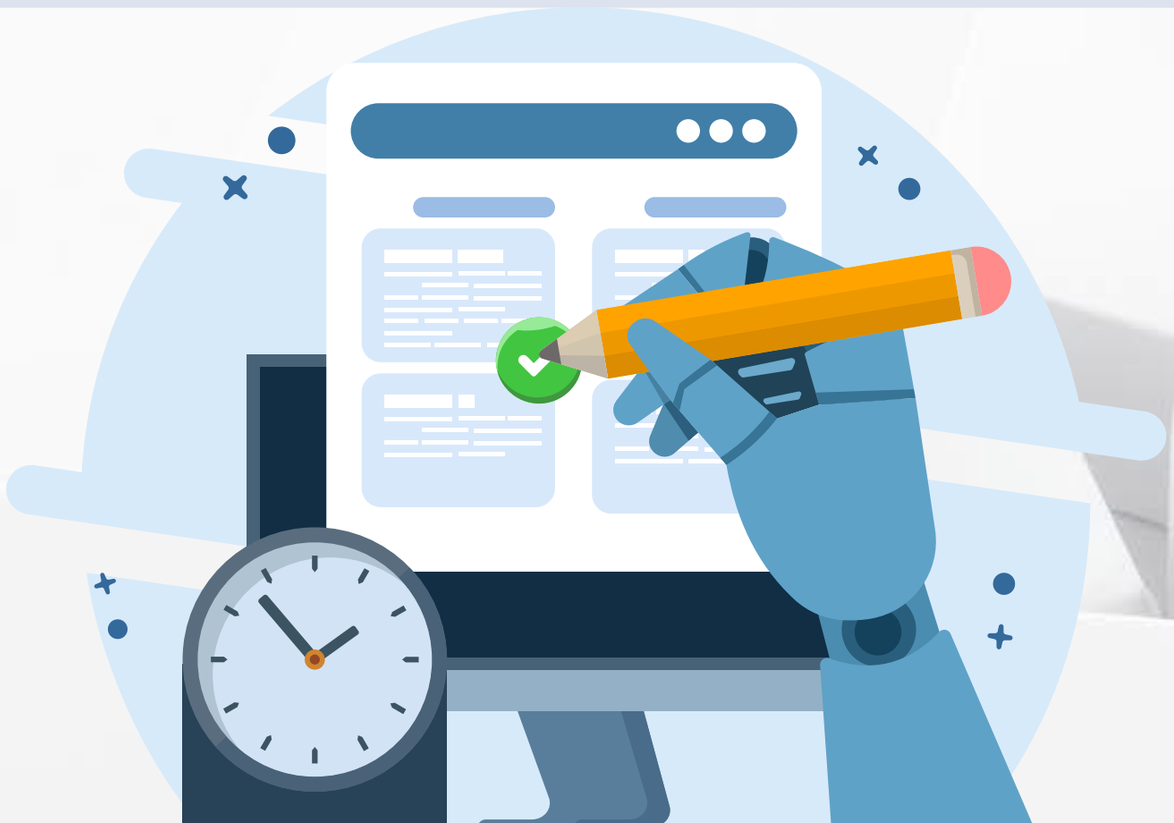
10 h

POR DEV. CADA SEMANA

Caso Instana

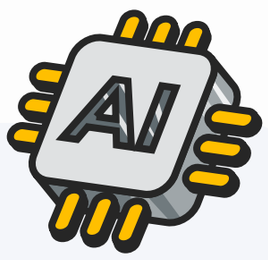
Ingeniería

El mapeo automático de dependencias y las tareas automatizadas seleccionadas generan un ahorro de 10 horas semanales por desarrollador.



MOTOR 03 · CONECTAR

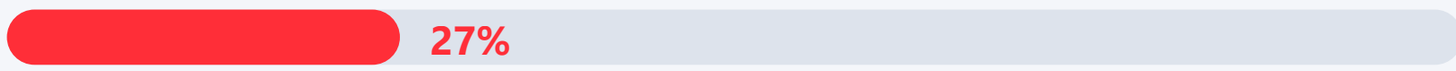
Darle manos a la IA sin descuidar la caja fuerte



957

APLICACIONES PROMEDIO

Solo el 27% están conectadas de manera efectiva



Fuente: MuleSoft

Si vas a darle capacidad de acción a un agente sobre tus sistemas legados (por ejemplo, pedir que congele una cuenta), necesitas fronteras blindadas.

LA REGLA ES SIMPLE

La IA propone; la plataforma autoriza, limita y registra.

Entre el agente de IA y los sistemas core legados se instala un AI Gateway (como IBM webMethods o IBM API Connect) que opera en tres capas estrictas:



Agente de IA

↓

CAPA 1

Model Context Protocol (MCP)

Convierte y expone las APIs REST del negocio como herramientas gobernadas que el agente puede descubrir e invocar.

CAPA 2

Filtros activos

Ejecuta enmascaramiento de información personal identificable (PII) al vuelo para evitar fugas de datos sensibles.

CAPA 3

Validación estricta

Exige cumplimiento con OAuth 2.0, FAPI 2.0, autenticación mutua mTLS obligatoria para ecosistemas abiertos y control de tráfico mediante rate-limiting.

↓

Sistemas core legados



MOTOR 04 · OPERAR

Infraestructura curada por catálogo, cero inventos

Dejar que una IA cree servidores por su cuenta en plena temporada alta es buscarse un dolor de cabeza financiero y operativo. La infraestructura tiene que ser inmutable y declarada desde catálogos curados, nunca por autonomía salvaje.



Flujo guiado por intención (Intent-Driven Workflow)

con Terraform y Red Hat OpenShift

1 La petición

Un operador solicita un entorno usando lenguaje natural (Intent).

2 La evaluación inteligente

El Terraform MCP Server, apoyado en la visibilidad casi en tiempo real de Infragraph, revisa el catálogo en HCL preaprobado sin inventar recursos.

App Servers · Database Clusters · Load Balancers · Network

3 Los frenos automáticos

Antes de tocar un solo recurso, el sistema detecta desviaciones de configuración (Drift) y aplica reglas de presupuesto (FinOps) junto con Policy as Code.

4 La ejecución inmutable (Día 2)

Si pasa las validaciones, se ejecuta terraform apply o Terraform Actions aprobadas.

TERRAFORM

```
resource 'terraform_action' 'rotate_keys' {
  action = 'execute'
  policy = 'approved'
}
```

-80% **Impacto comprobado: el caso CarGurus redujo un 80% los tiempos de aprovisionamiento (lead time) con catálogos empresariales de HCP Terraform.**

CAPÍTULO 03

Gobierno desde el día cero

La letra menuda que te salva el pellejo.

03

Existe el mito de que meter seguridad y gobierno vuelve lentos a los equipos. Los datos estructurales del estudio IBM IBV 2026 demuestran exactamente lo contrario: automatizar barreras de contención (guardrails) mediante policy as code desde el Día Cero es la licencia para experimentar y desplegar rápido sin romper producción.

Comparadas con las empresas que siguen haciendo revisiones manuales, las organizaciones con gobernanza embebida logran:



16x

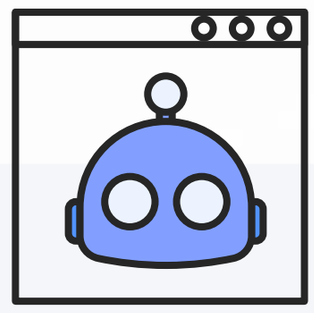
más agentes desplegados en producción masiva.

4x

menos presupuesto consumido por proyecto (operan con apenas 1/4 del capital histórico).

PRINCIPIOS

Los 4 principios innegociables de gobierno



01

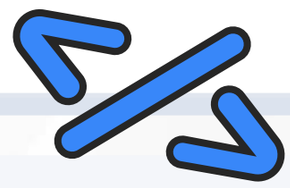
Cero agentes que “solo contestan”

Todo proyecto debe nacer con arquitectura sólida, justificación de negocio y métricas claras desde el Día Cero.

02

Transparencia y clasificación

Tipificación obligatoria de cada proyecto según su nivel de autonomía e impacto operativo.



03

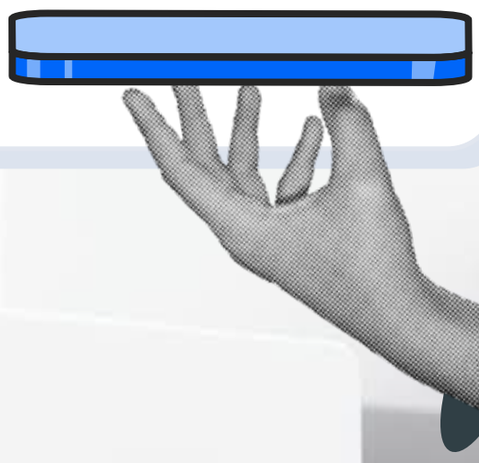
Confianza Cero (Zero Trust)

Redes privadas e identidad máquina diseñadas desde la base, no como parches de última hora.

04

Calidad verificable

Pruebas automatizadas rigurosas, ejercicios de Red Teaming y observabilidad extrema en producción.





SEGURIDAD

Estándares Zero Trust e identidad máquina

Dentro de una arquitectura blindada (conectando el agente de IA, la base de datos vectorial Vector DB y las APIs externas), aplican tres reglas duras:

Aislamiento de red total

Prohibido salir a internet público. Todo corre dentro de una VPC aislada usando AWS PrivateLink o Azure Private Endpoint.

Gestión estricta de secretos

Prohibición absoluta de API Keys estáticas. Se usa Vault para generar tokens dinámicos efímeros con tiempo de vida (TTL) corto y autenticación bidireccional mTLS.

Control de salida (Egress Filtering)

Filtra el tráfico saliente para bloquear cualquier intento de exfiltración de datos provocado por ataques de Prompt Injection.

AUDITORÍA

Observabilidad y evidencia inmutable (WORM)

Para pasar cualquier auditoría de cierre de año sin sudar frío, el sistema exige umbrales técnicos estrictos y trazabilidad imposible de manipular.

MÉTRICAS DE CALIDAD



< 5 s

Latencia p95



≥ 95%

Precisión (Accuracy)



< 5%

Tasa de alucinación

Respaldadas por un conjunto de **50 a 500+ Golden Examples**

Caja negra inalterable

Todo el registro de auditoría (Audit Trail / SIEM) se guarda en almacenamiento WORM (Write-Once-Read-Many). Cada traza exige formato JSON estructurado y firma criptográfica para garantizar que nadie altere la evidencia:

JSON · TRAZA FIRMADA

```
{
  "Trace_ID": "x98b2",
  "IAM_Identity": "agent_role_01",
  "Input": "[hash_encrypted]",
  "Crypto_Signature": "verified"
}
```



CAPÍTULO 04

De la teoría a la cancha: el caso real de Banco Demo

04

Originación de producto regulado diseñada por ASSIST

Para aterrizar todo este engranaje, miremos un caso real de originación de producto regulado diseñado por ASSIST (IBM Gold Partner): Banco Demo v1.1.0 (DemoCréditos y Fraude), desplegado en la IBM Technology Zone.

TIEMPO DE SALIDA A PRODUCCIÓN



EL RETO DEL NEGOCIO

“¿Se puede aprobar un crédito de COP 20 millones a 36 meses para el cliente DEMO-001?”

El agente debe consultar su perfil, evaluar el riesgo y explicar su decisión usando exclusivamente las herramientas gobernadas del banco.



RECORRIDO A

Construcción y ejecución

La arquitectura separa limpiamente cómo se construye la aplicación (lo que ya existe y corre en línea sólida) de cómo la consume el agente de forma gobernada (la integración MCP y gateway propuesta en línea discontinua).

Línea sólida: lo que ya existe y corre

DESARROLLO

1 **IBM Bob**

Acelera el código, la interfaz, las pruebas y genera el contrato OpenAPI (revisado por el equipo).

↓

AUTOMATIZACIÓN

2 **Terraform**

Versiona y reproduce la infraestructura como código con provisión demostrable.

↓

PLATAFORMA

3 **Red Hat OpenShift**

Hospeda la aplicación contenedorizada y publica su ruta operativa.

↓

DESPLEGADO

4 **Banco Demo**

Expone los servicios REST conservando las reglas de negocio.

— RECORRIDO B

Herramientas y agente

Línea discontinua: integración MCP y gateway propuesta

AGENTE

1

OpenAI Codex

Actúa como cliente MCP propuesto; descubre e invoca herramientas según permisos y explica la decisión sobre resultados reales del backend.

GATEWAY

2

IBM API Connect / webMethods

Aplica autenticación, políticas de autorización, límites y observabilidad. IBM webMethods aporta la capa de integración empresarial según el entorno y la versión disponibles.

HERRAMIENTAS

3

Servidor MCP

Generado directamente a partir de los esquemas y operaciones del contrato OpenAPI + REST.

BACKEND

4

APIs del banco

Ejecutan la lógica central y devuelven las respuestas.

HERRAMIENTAS GOBERNADAS

Las 5 únicas herramientas que el agente puede tocar

El agente no navega a ciegas ni tiene acceso libre a las bases de datos; el servidor MCP le expone exactamente cinco operaciones:

01	list_demo_customers	MCP
02	get_customer_credit_profile	MCP
03	calculate_demo_credit_score	MCP
04	check_demo_fraud_signals	MCP
05	evaluate_demo_loan_application	MCP

Cómo se engranan los 4 motores (Hoy vs. Evolución)

		EN LA DEMO (HOY)	EN LA EVOLUCIÓN
01	**Desarrollo asistido de software**	IBM Bob construyó el banco y su contrato OpenAPI.	El patrón se extiende a nuevas soluciones e integraciones.
02	**Gestión de datos en tiempo real**	Las APIs operan sobre información sintética y respuestas bajo demanda.	Se conectan fuentes transaccionales y eventos operativos en vivo.
03	**Integración conectada con acción**	Se reutilizan las APIs existentes con MCP y API Connect propuestos.	Herramientas con permisos granulares, políticas y trazas completas.
04	**Infraestructura inmutable automatizada**	La aplicación corre en OpenShift con Terraform dentro del recorrido.	Se estandarizan imágenes, despliegues y configuración externa.

NOTA DE TRANSPARENCIA

En esta demostración de crédito y fraude los resultados son exclusivamente simulados; no se consultan burós reales, no existen confirmaciones reales de fraude ni se realizan desembolsos de dinero.

HERRAMIENTAS GOBERNADAS

Las 5 únicas herramientas que el agente puede tocar

El agente no navega a ciegas ni tiene acceso libre a las bases de datos; el servidor MCP le expone exactamente cinco operaciones:

1	`list_demo_customers`	MCP
2	`get_customer_credit_profile`	MCP
3	`calculate_demo_credit_score`	MCP
4	`check_demo_fraud_signals`	MCP
5	`evaluate_demo_loan_application`	MCP



NOTA DE TRANSPARENCIA, SIN LETRA MENUDA

En esta demostración de crédito y fraude los resultados son exclusivamente simulados; no se consultan burós reales, no existen confirmaciones reales de fraude ni se realizan desembolsos de dinero.



HOY VS. EVOLUCIÓN

Cómo se engranan los 4 motores

01 Desarrollo asistido de software

EN LA DEMO (HOY)

IBM Bob construyó el banco y su contrato OpenAPI.

EN LA EVOLUCIÓN

Este patrón se extiende a nuevas soluciones e integraciones.

02 Gestión de datos en tiempo real

EN LA DEMO (HOY)

Las APIs operan sobre información sintética y respuestas bajo demanda.

EN LA EVOLUCIÓN

Se conectan fuentes transaccionales y eventos operativos en vivo.

03 Integración conectada con acción

EN LA DEMO (HOY)

Se reutilizan las APIs existentes con MCP y API Connect propuestos.

EN LA EVOLUCIÓN

Se exponen herramientas con permisos granulares, políticas y trazas completas.

04 Infraestructura inmutable automatizada

EN LA DEMO (HOY)

La aplicación corre en OpenShift con Terraform dentro del recorrido.

EN LA EVOLUCIÓN

Se estandarizan imágenes, despliegues y configuración externa.

ASSIST

ANTES DE ACTIVAR EL MODO VACACIONES

Cierra el año con la infraestructura blindada

¿Listo para industrializar tu IA?

AGENDA TU SESIÓN →

