Bits AI SRE & LLM Observability

Evita que las alertas se conviertan en incidentes con RCA autónoma

Monitorea, soluciona problemas, asegura e itera sobre aplicaciones LLM y agentes de IA

Octubre 8, 2025

epidata <επι/> deploying ideas

DATADOG

Gustavo Souza

Partner Solutions Architect
gustavo.souza@datadoghq.com

https://www.linkedin.com/in/gpsouza93/

Vuestro Orador

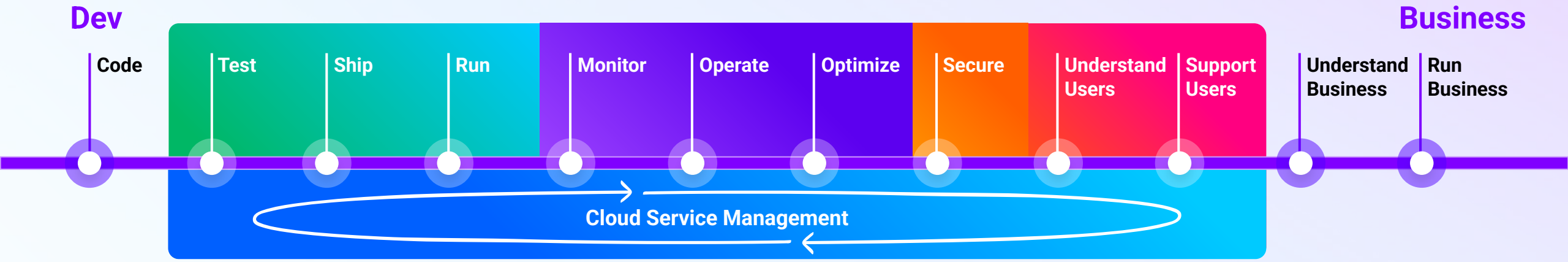
epidata
deploying ideas

DATADOG

Datadog es la **plataforma única** para todas sus **necesidades de observabilidad**

epidata deploying ideas

Datadog Platform



Software Delivery

- CI Visibility
- Test Optimization
- Continuous Testing

Monitor & Operate

- Infra Monitoring
- Network Monitoring
- APM
- Synthetics
- Log Mgmt
- Universal Service Monitoring
- Observability Pipelines
- LLM Observability

Optimize

- Continuous Profiler
- Cloud Cost Mgmt
- Database Monitoring
- Data Streams Monitoring
- Data Jobs Monitoring

Secure

- Cloud Security
- Code Security
- Cloud SIEM
- Sensitive Data Scanner
- Workload Protection
- App and API Protection

Analyze

- RUM
- Heatmap/ Clickmap/ Scrollmap
- Mobile App Testing
- Session Replay

Cloud Service Management

- Incident Management
- On-Call
- Workflow Automation
- App Builder
- Software Catalog
- Resource Catalog
- Case Management
- Event Management

Applied AI

Natural Language Querying • Root Cause Analysis • Anomaly Detection • Impact Analysis • Proactive Alerts • Autonomous Investigations • Bits AI

Demos un paso atrás… ¿Qué ofrece Datadog?

Capacidades de IA en toda la plataforma Datadog

IA aplicada

- Capacidades embebidas en toda la plataforma: Detección proactiva de anomalías, análisis de causa raíz (antes **Watchdog**), consulta por lenguaje natural, explicaciones de gráficos, detección de información confidencial y el **Agente Autónomo de Bits** (incluye Action Interface)

Herramientas para que los clientes desarrollen y monitoreen su propia IA

Observabilidad de IA

- **Observabilidad de LLMs & Agentes**: monitoriza el rendimiento, calidad y seguridad por todo el ciclo vital de los prompts
- **Integraciones pre construidas** para monitorizar tecnologías comunes en proyectos de IA (OpenAI, AWS Bedrock, Gemini)
- **GPU Monitoring:** monitoriza la infraestructura de IA
- **MCP Server:** acceda a datos de Datadog para sus propios agentes

¿Cuanto tiempo tus equipos gastan **respondiendo a alertas**?

epidata
deploying ideas

El costo de una respuesta de alerta lenta

Horas de ingeniería

Confianza del cliente

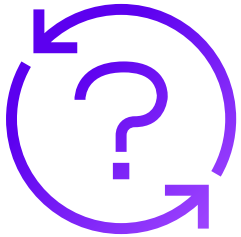
Pérdida de ingresos

$23.8k

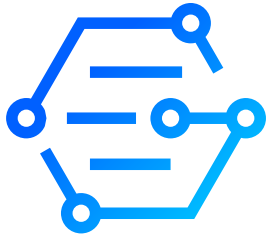
por minuto es el costo de la indisponibilidad no planificada para grandes empresas, según una investigación de la industria de TI.

https://www.bigpanda.io/blog/it-outage-costs-2024/

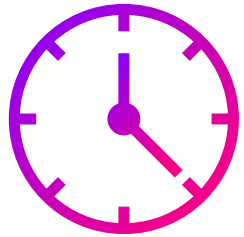
La respuesta ante alertas tiene limitaciones humanas



Triaje
lento

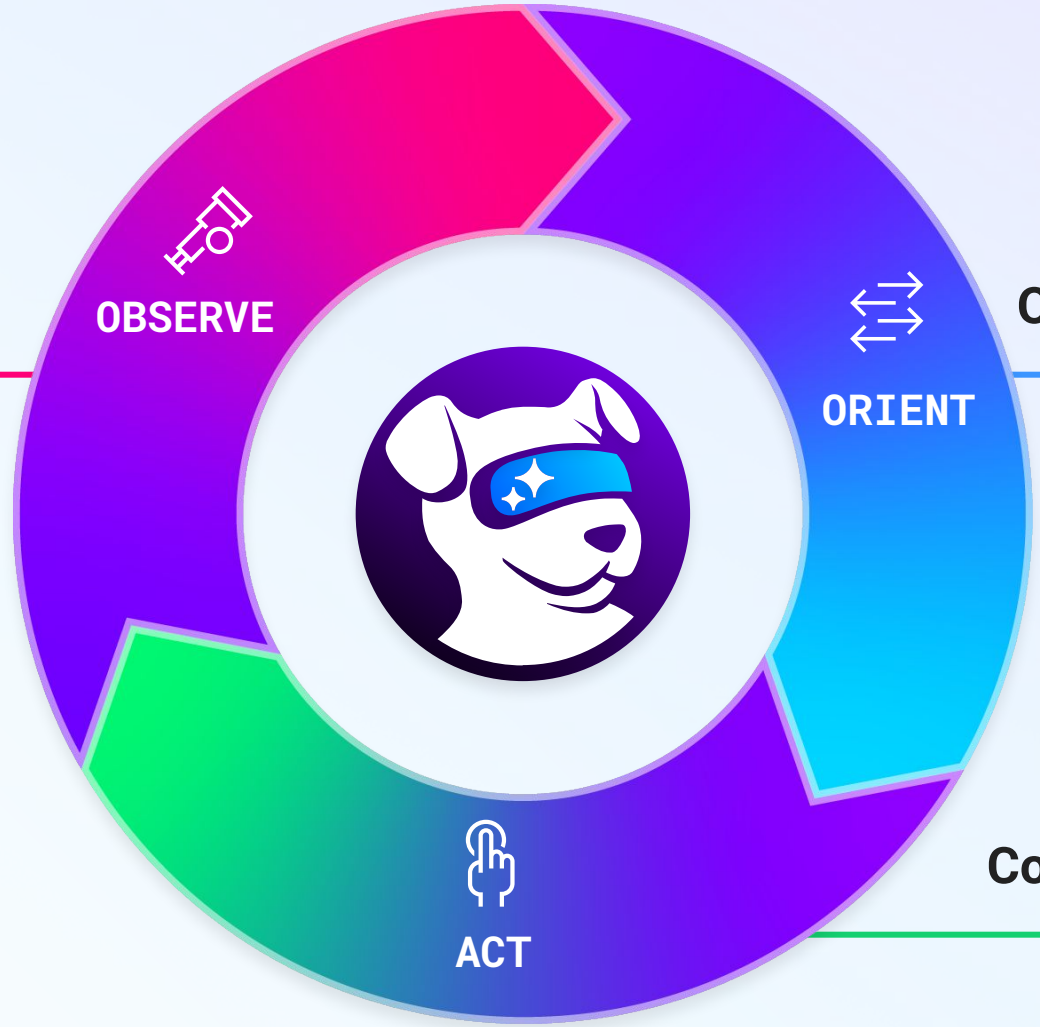


Procesos
manuales



Soluciones
retrasadas

Introduciendo el Agente Autónomo de Bits



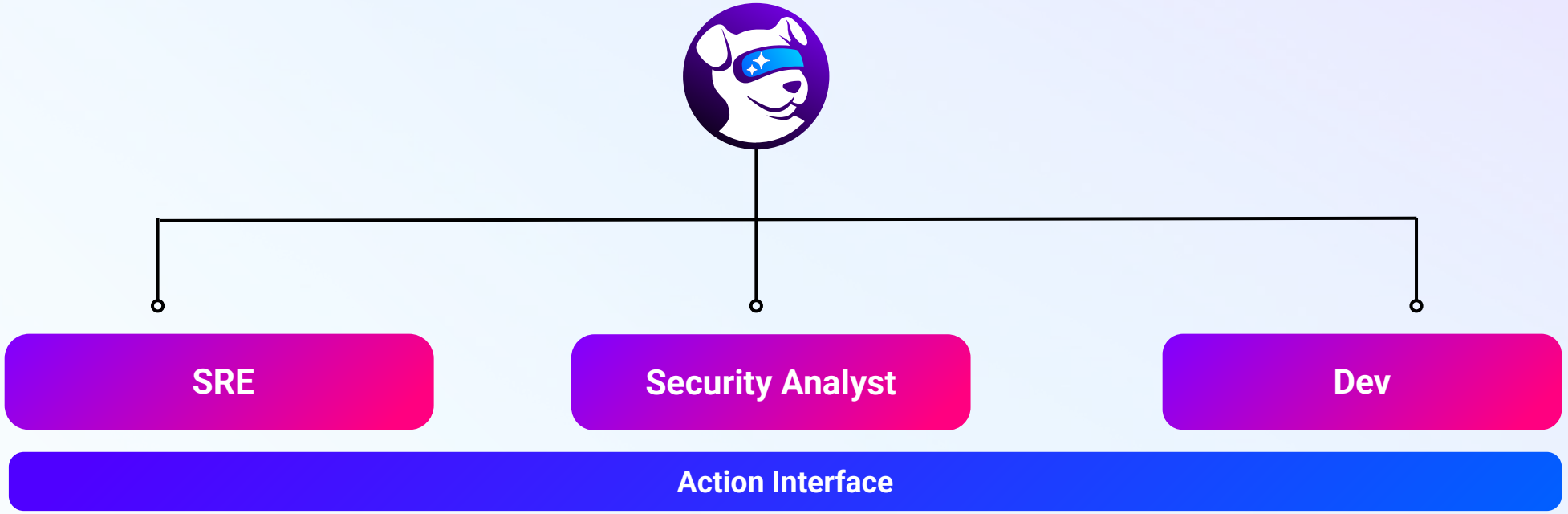
Operacionalmente consciente

Organizacionalmente consciente

Contextualmente consciente

** Los agentes de IA no son lo mismo que el agente de Datadog*

Empezando por las 3 funciones principales del ciclo vital de DevSecOps



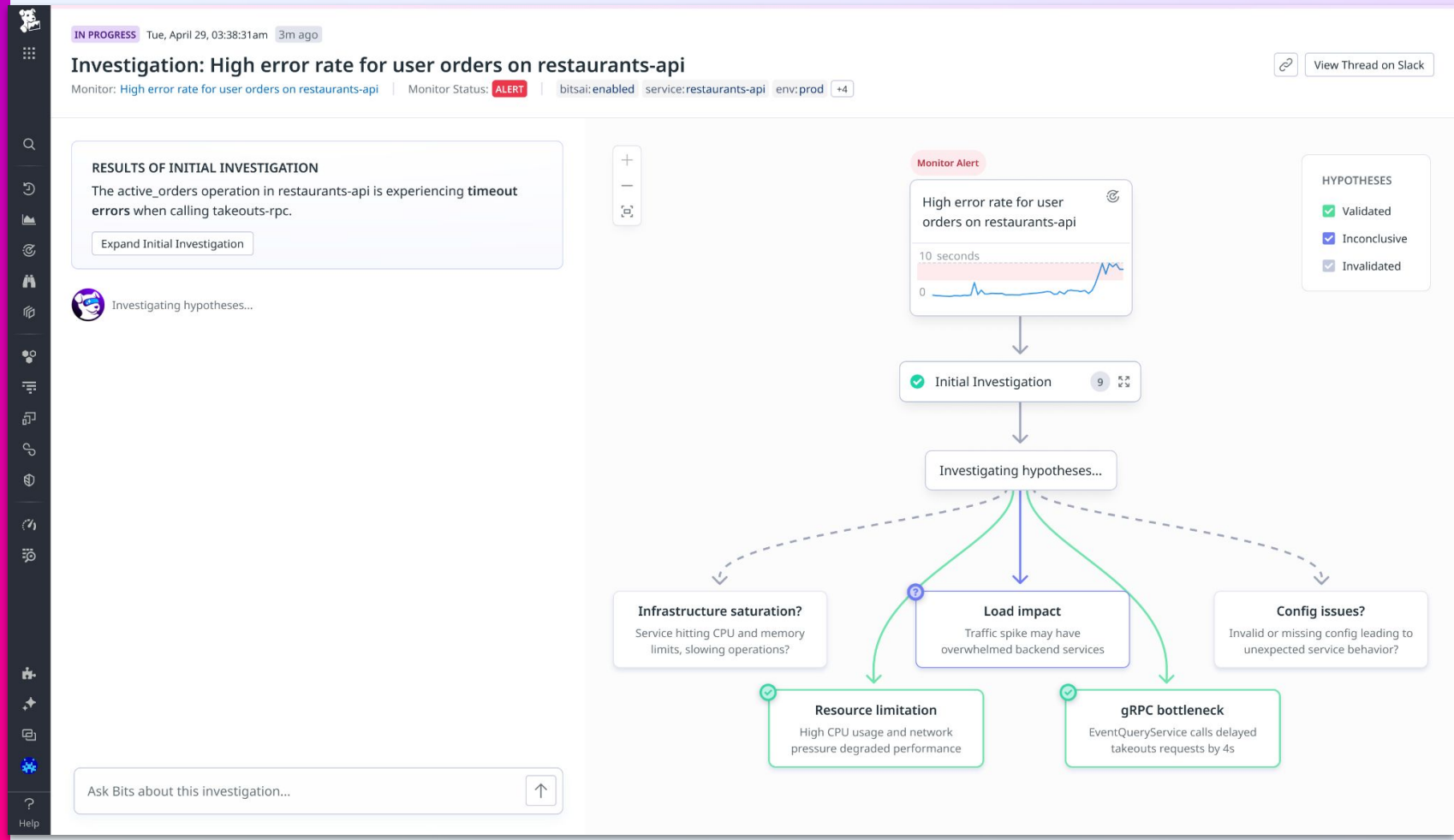
Bits AI *SRE*

LA at DASH

DATADOG

On-call 24/7, investigando alertas y coordinando incidentes en escala

- Investiga la causa raíz de por detrás de cada alerta de forma autónoma - antes de que llegues a tu laptop
- Genera actualizaciones en tiempo real, indica los incidentes relacionados y ayuda con tareas post incidente



Usuarios que lo amarán:

- SREs
- DevOps
- Ingenieros de Plataforma

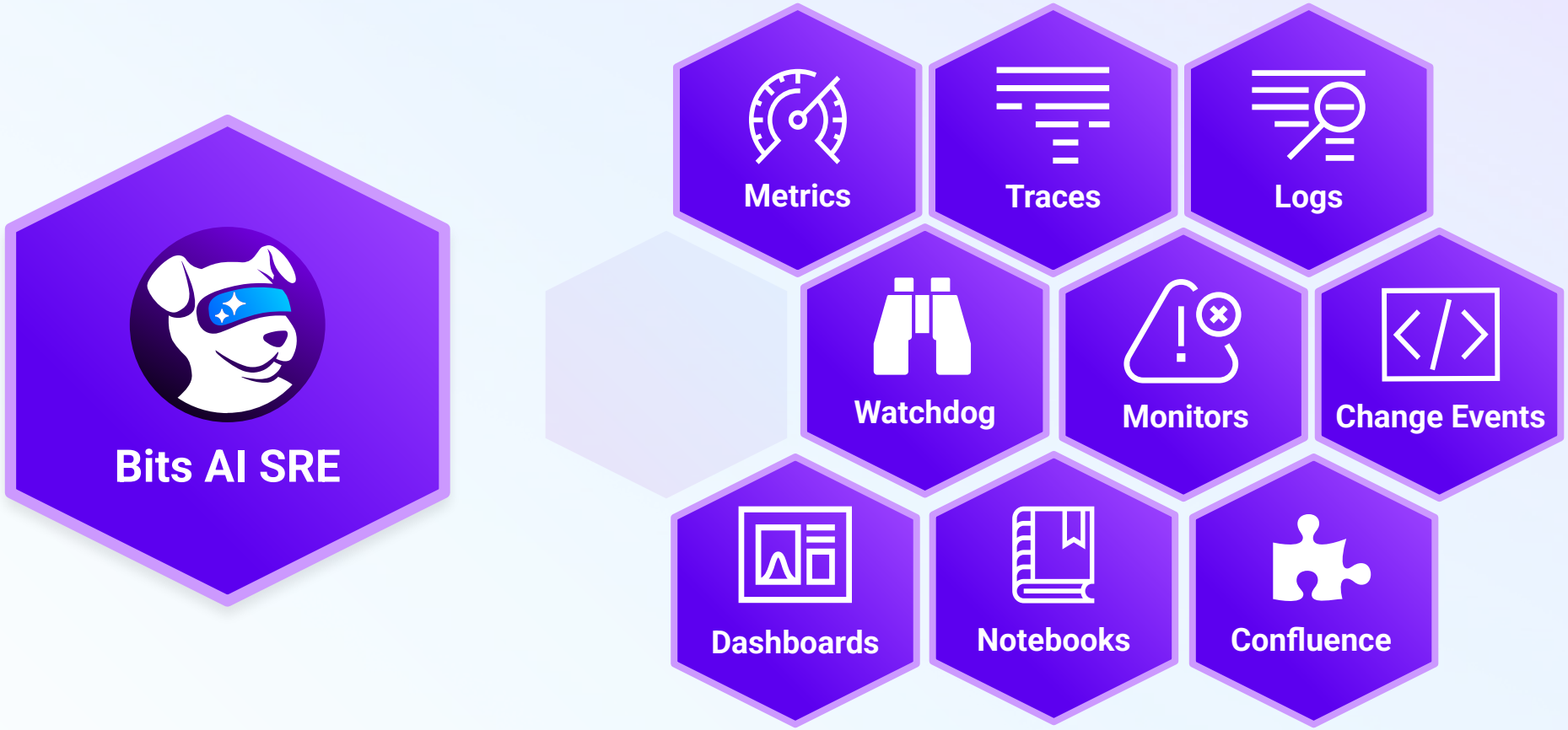
👇 Empieza por aquí

- Regístrate para disponibilidad limitada

📓 Más información

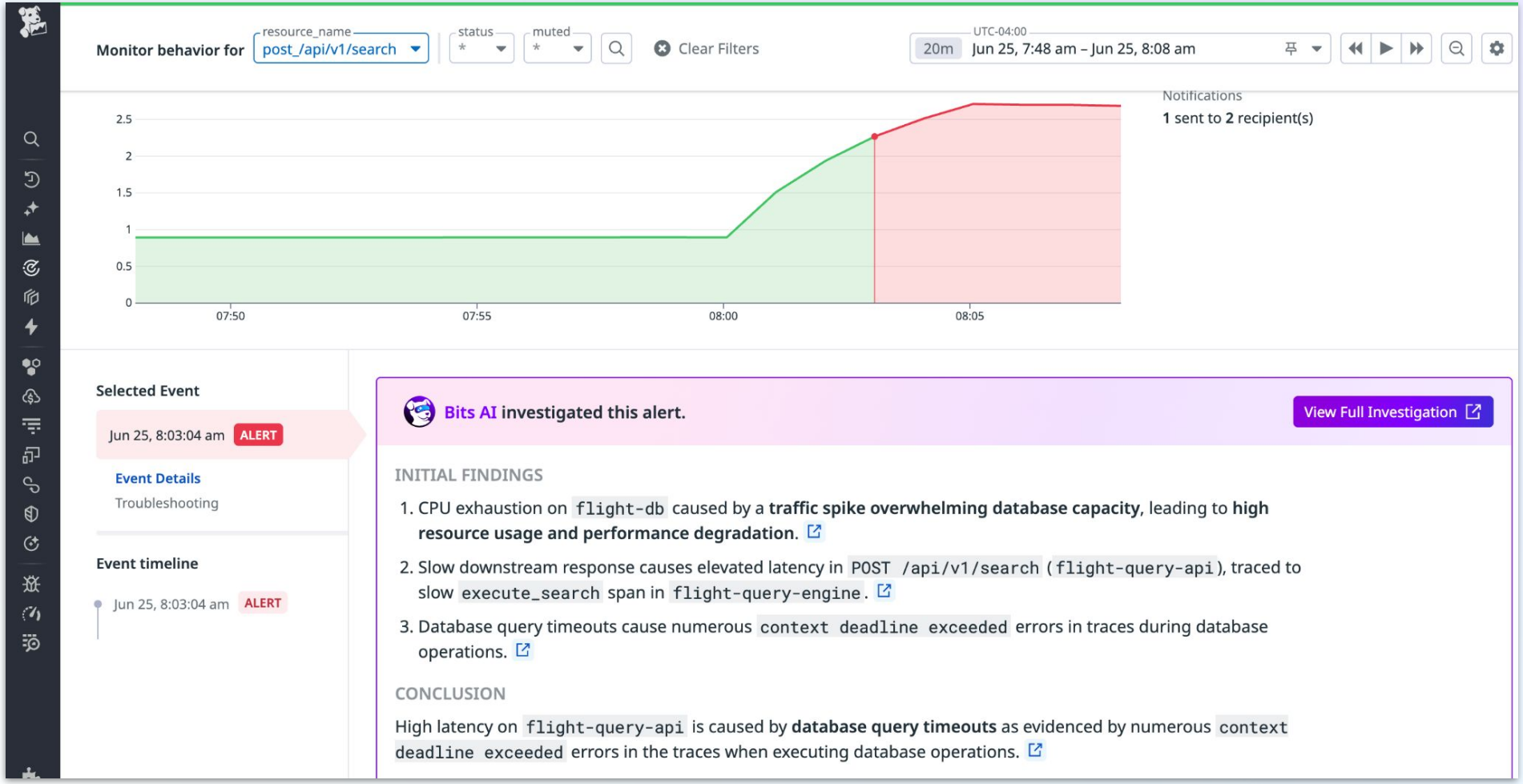
- Presentando Bits AI SRE, tu compañero de IA on-call

Bits AI SRE utiliza herramientas y telemetría de Datadog para investigar alertas de forma autónoma



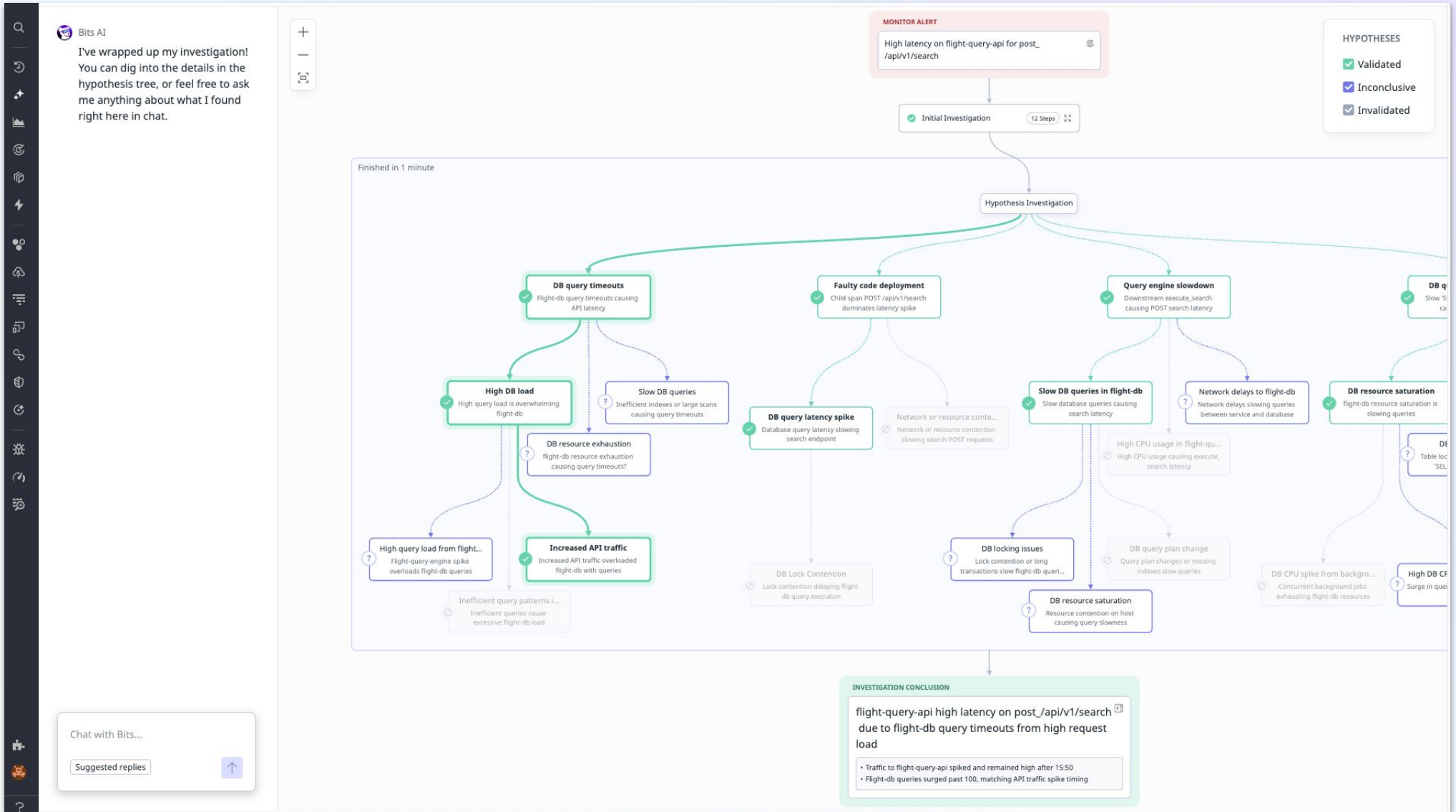
Revisa un resumen del alerta minutos después que se activa

Bits AI SRE entrega sus hallazgos directamente en el monitor que se activó para que pueda tomar medidas inmediatas



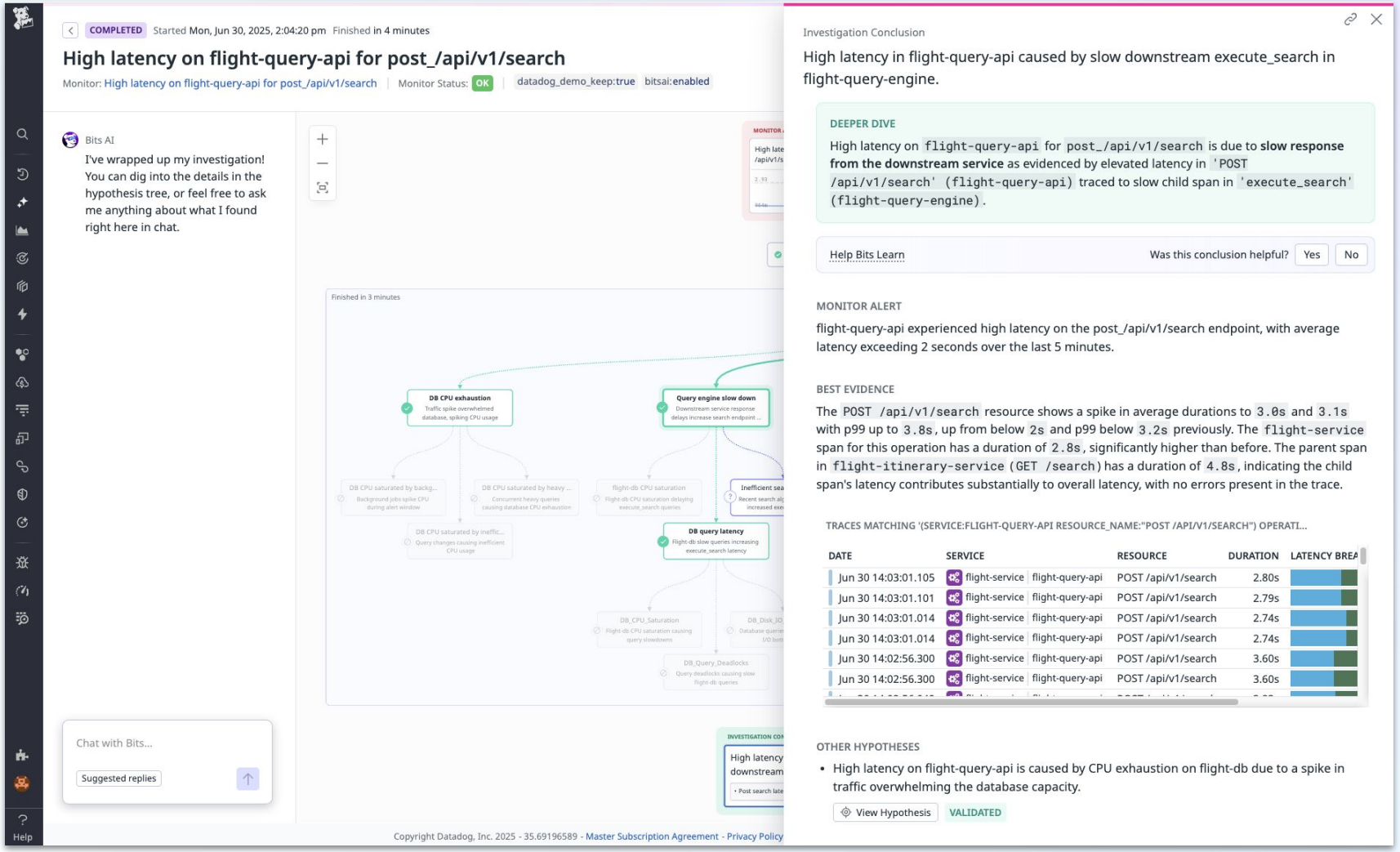
Investiga múltiples causas en escala usando Datadog

Bits AI SRE razona de forma autónoma utilizando herramientas Datadog en cada paso de su investigación



Descubra la causa raíz más rápidamente y evite incidentes futuros

Bits AI SRE aprende por sí solo y se le puede pedir que recuerde o mejore los pasos en su evaluación



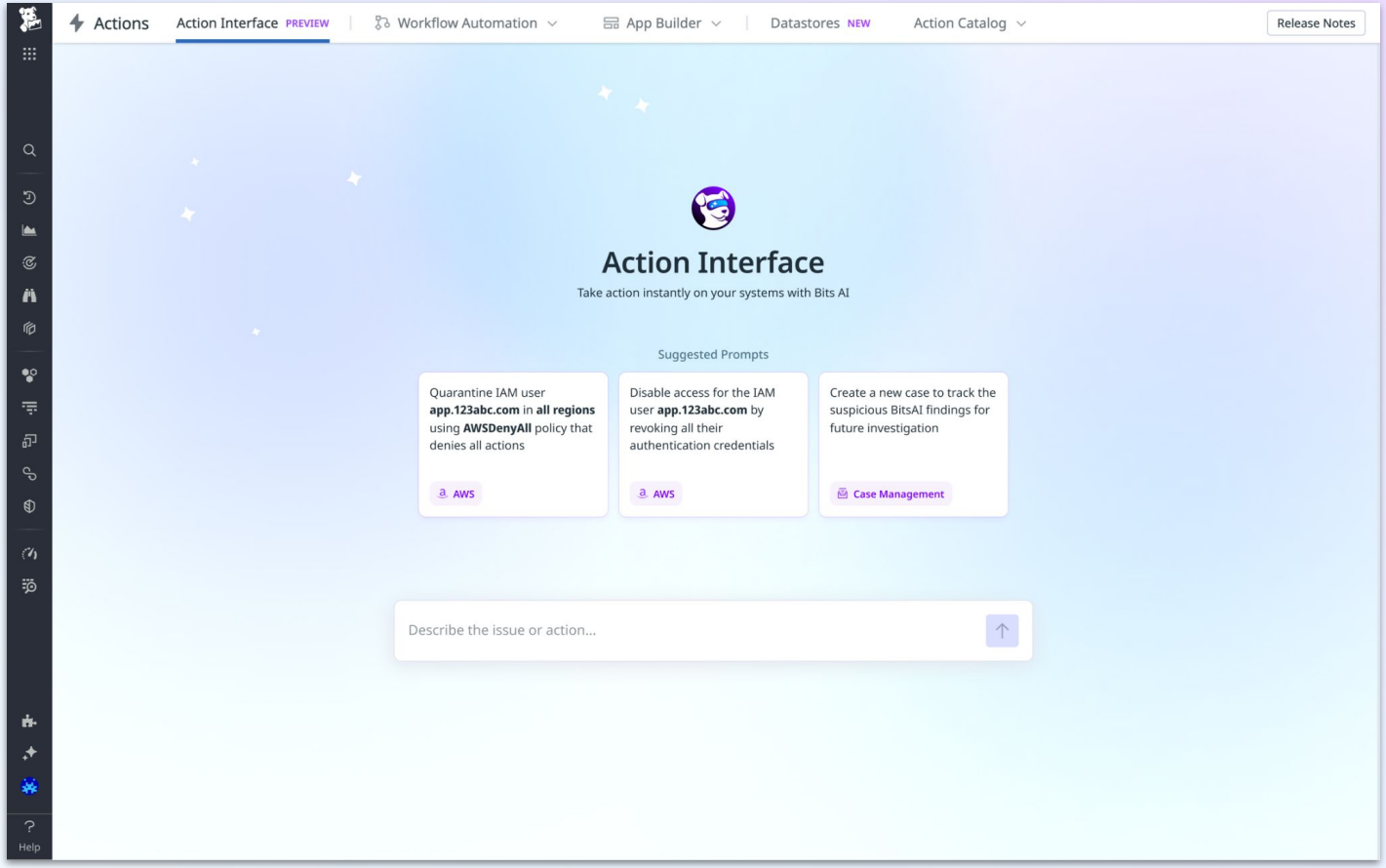
Preview at DASH

Action Interface

Incluido con Workflows y App Builder

Toma acción a partir de los descubrimientos en Datadog sin tener que crear aplicativos o flujos de trabajo

- Describa su tarea usando lenguaje natural y Action Interface te sugiere una acción correspondiente del catálogo de acciones
- Aprobación con un click desencadena el comando en tu infraestructura



Usuarios que lo amarán:

- SREs
- DevOps
- Ingenieros de Plataforma
- Ingenieros de Seguridad

🎯 Ideal para clientes que...

Desean tomar medidas desde Datadog, pero no tienen aplicaciones ni flujos de trabajo.

👇 Empieza por aquí

Formulario para acceso

ANTES

Recursos desperdiciados y costosos tiempos de inactividad

Tiempos de respuesta lentos

Resolución de problemas manual

Investigaciones prolongadas

DESPUÉS

Investigaciones de IA, completadas en minutos

Investigaciones inmediatas

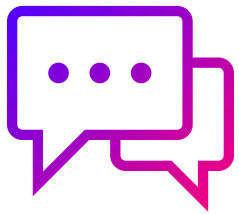
Análisis de causa raíz autónomo

Resoluciones más rápidas

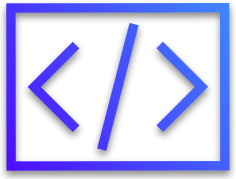
Clientes construyendo sus propias IAs

epidata <επι/> deploying ideas

Varios casos de uso que surgen con los LLMs



Chatbots



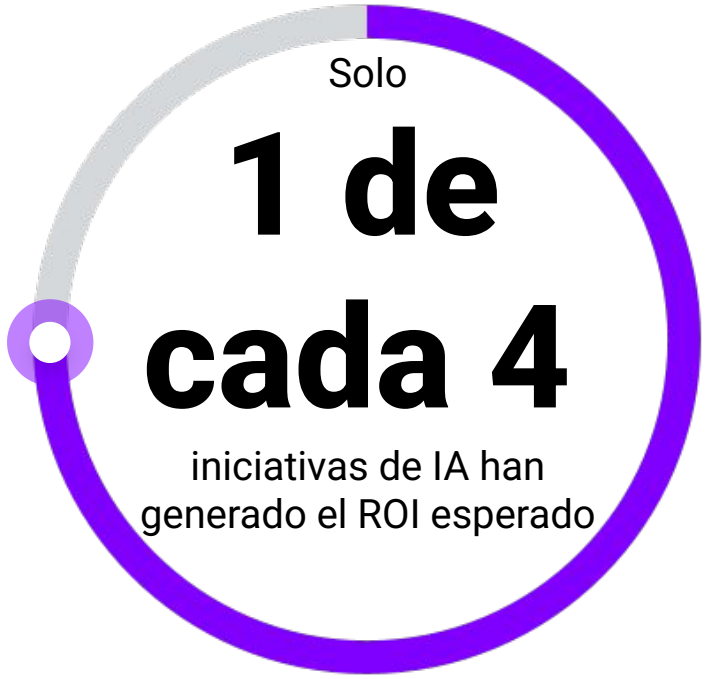
Generación de Código

Resumen & Traducción

Generación de Contenido, etc.

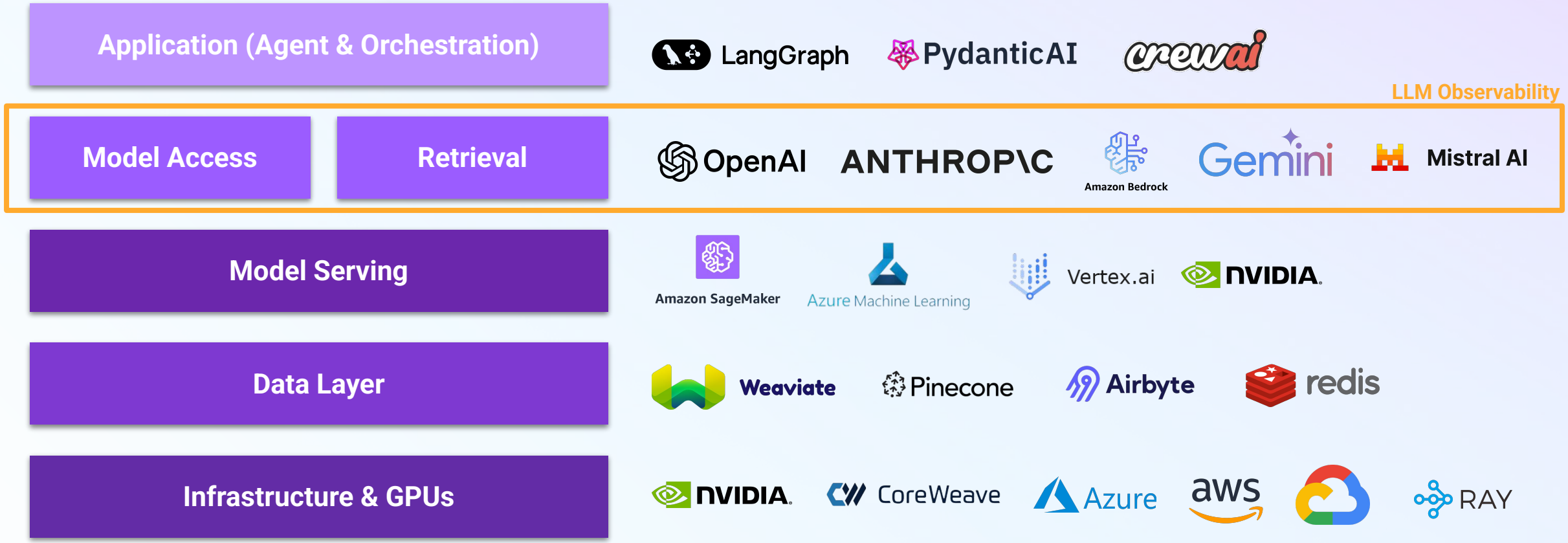
La Importancia de la Observabilidad de LLM

A lo largo de los últimos 3 años, CEOs dicen:



Fuente: IBM's 2025 CEO Study (link)

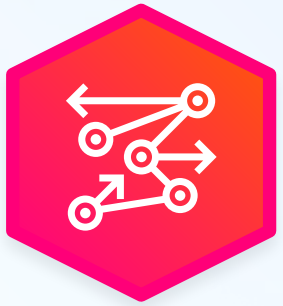
Se requiere a los desarrolladores navegar con un nuevo y completo conjunto de tecnologías



Las aplicaciones LLM en producción introducen nuevos problemas

Complejidad	Los flujos de LLMs y agentes son distribuidos, no deterministas y, a veces, autónomos
Alucinaciones & Calidad	Las respuestas inexactas requieren una evaluación y corrección continuas para garantizar la confiabilidad
Costos de LLM	El uso de tokens puede salirse de control, especialmente bajo cargas de trabajo de producción impredecibles
Seguridad & Protección	Las aplicaciones LLM son vulnerables a nuevos tipos de ataques, como el hacking del prompt y las filtraciones de datos confidenciales

Aplicaciones de IA son una caja negra y los equipos están volando ciegos



Visibilidad limitada en el rendimiento y costos

Los equipos no saben dónde ocurren problemas de latencia, errores o costos en las cadenas y operaciones de los agentes, lo que ralentiza la resolución de problemas.

Problemas de calidad y seguridad no detectados

Las alucinaciones, las respuestas inseguras, las inyecciones rápidas y las filtraciones de información confidencial resultan en una mala experiencia de usuario y una reputación negativa.

Impacto poco claro de los cambios de prompts y modelos

Los equipos envían cambios sin saber cómo afectarán el rendimiento, la calidad y el costo.

AI Stack Monitoring

GA

Experiencia de monitoreo unificada

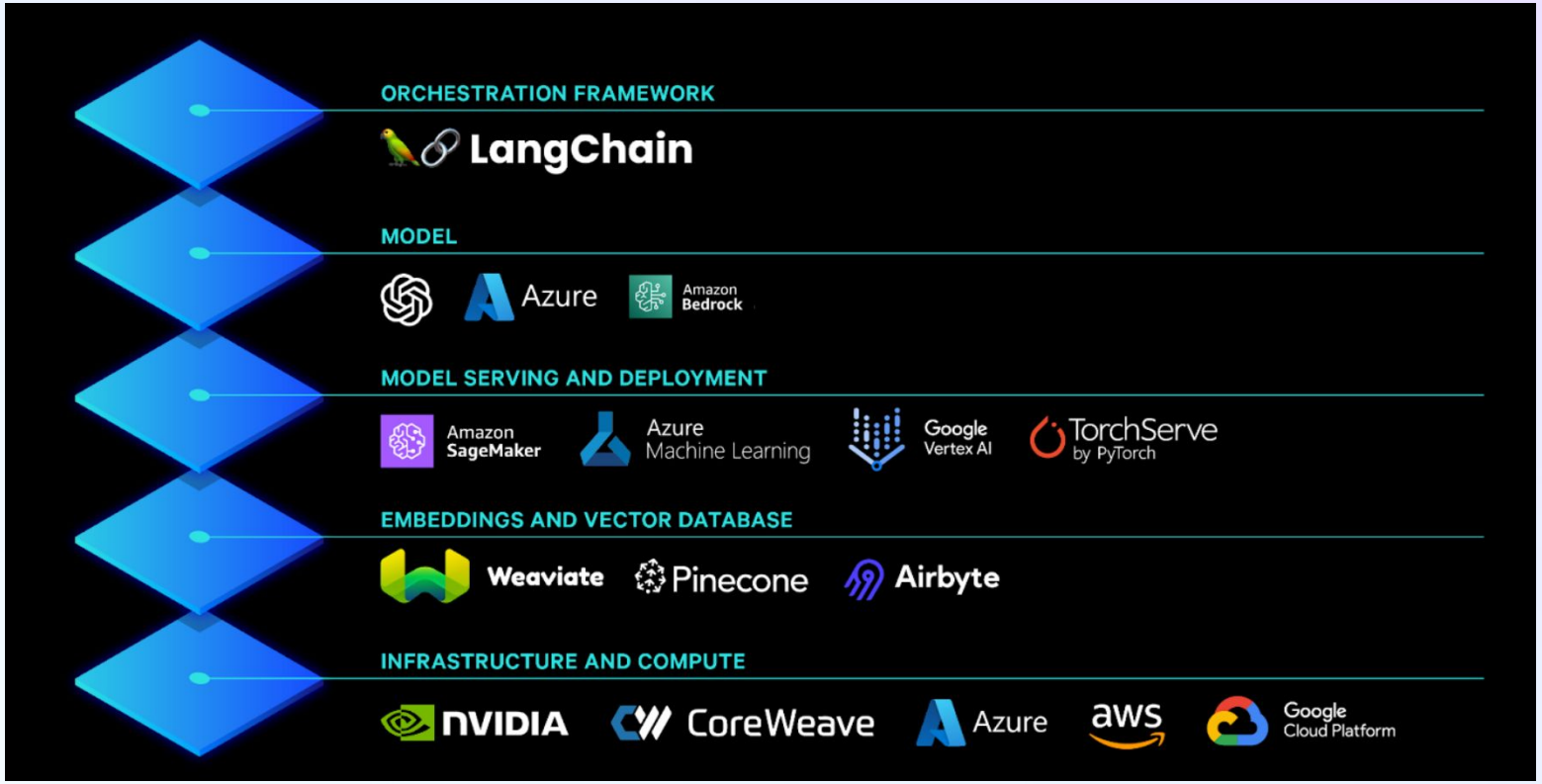
Amplio conjunto de integraciones para la observabilidad en diferentes capas de IA/ML, todo detrás de un único panel

Rápido tiempo de valorización

Dashboards y alertas prediseñados que encapsulan las mejores prácticas para monitorear estas tecnologías

Tecnologías emergentes

Las tecnologías más nuevas, como las bases de datos vectoriales, los frameworks de inferencia y la orquestación de IA, ahora están cubiertas con las integraciones de Datadog.



Usuarios que lo amarán:

- Desarrolladores
- SREs
- Ingenieros de aprendizaje automático

GPU Monitoring

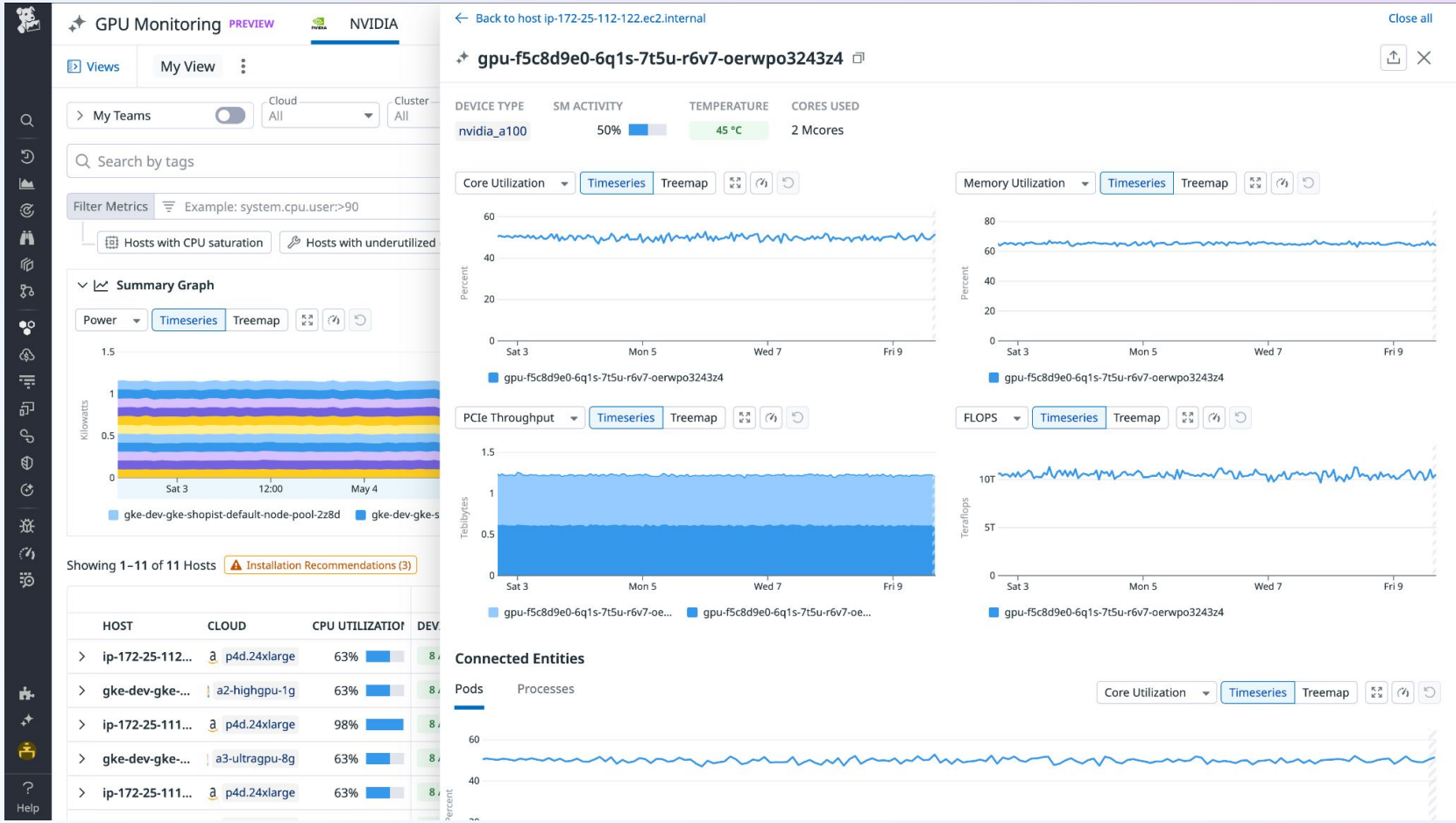
Nuevo SKU

GPU Monitoring le permite:

- **Mejorar la alocación de GPU** con la visibilidad del consumo a nivel de carga de trabajo (saturación del GPU, equipos subutilizados, calentamiento excesivo, etc.)
- **Evita gastos excesivos con GPU** al detallar el uso por job, pod y equipo
- **Soluciona latencia y fallas** más rápidamente al correlacionar métricas de la transferencia de datos y rendimiento de la infra

Usuarios que lo amarán:

- SREs
- Equipos DevOps
- Ingenieros de IA



🎯 La mejor opción para clientes que...

están buscando uso más eficiente y costo-efectivo del uso de GPUs que impulsan sus aplicaciones de IA y LLMs

👇 Empieza aquí

Lee el blog y regístrate para el preview

Core LLM Observability

GA

Monitoriza problemas operacionales & funcionales

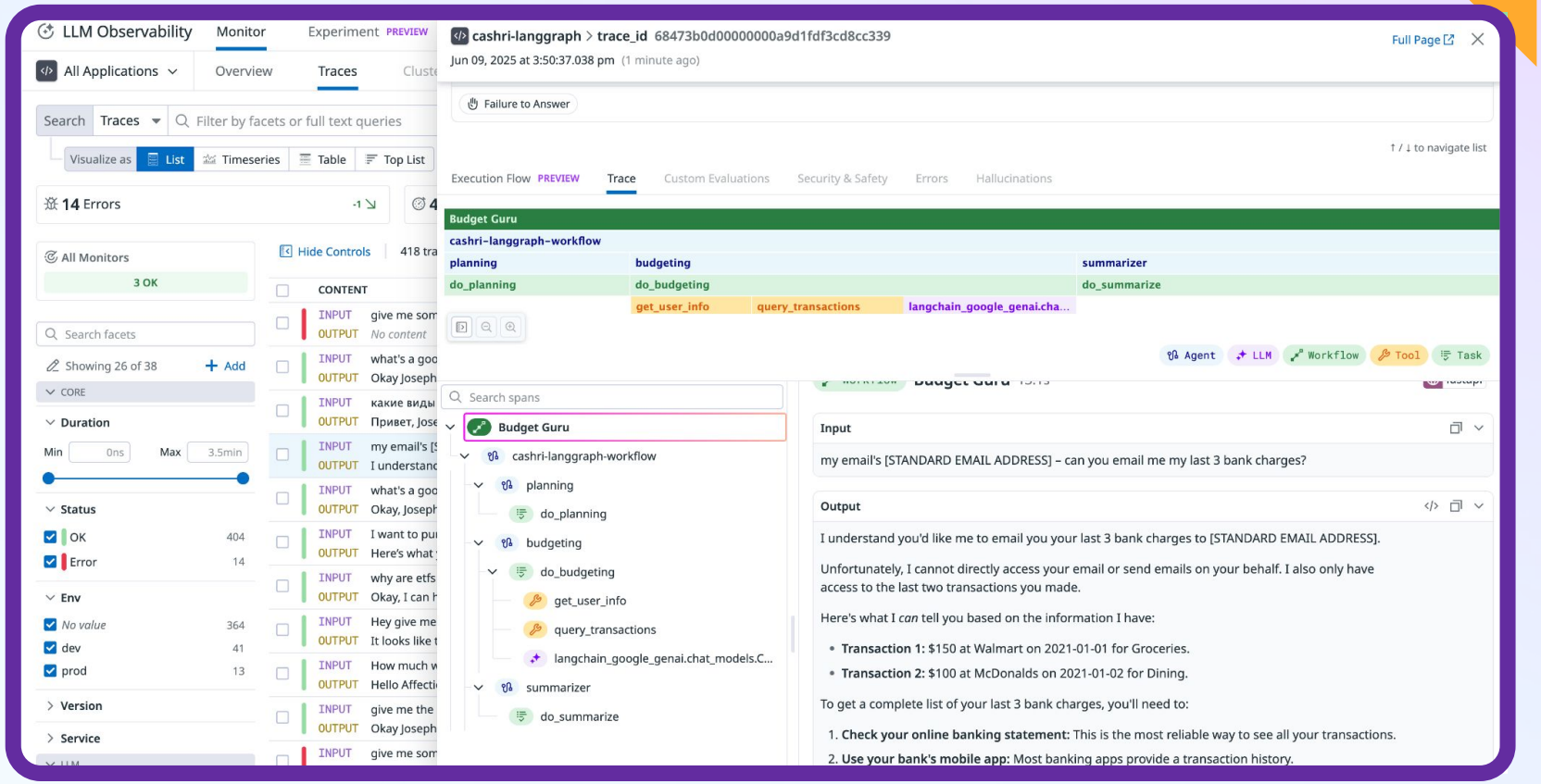
Detecta aumentos en errores, latencia y costos, y evalúa problemas de calidad y seguridad

Analiza problemas más rápidamente

Visualiza e identifica problemas en las operaciones de agentes y cadenas LLM paso a paso

Itera con rapidez y confianza

Genere conjuntos de datos y compare experimentos con indicaciones, modelos y cambios lógicos para mejorar su aplicación de LLM



Usuarios que lo amarán:

- Desarrolladores
- Ingenieros de aprendizaje automático
- Cientistas de datos

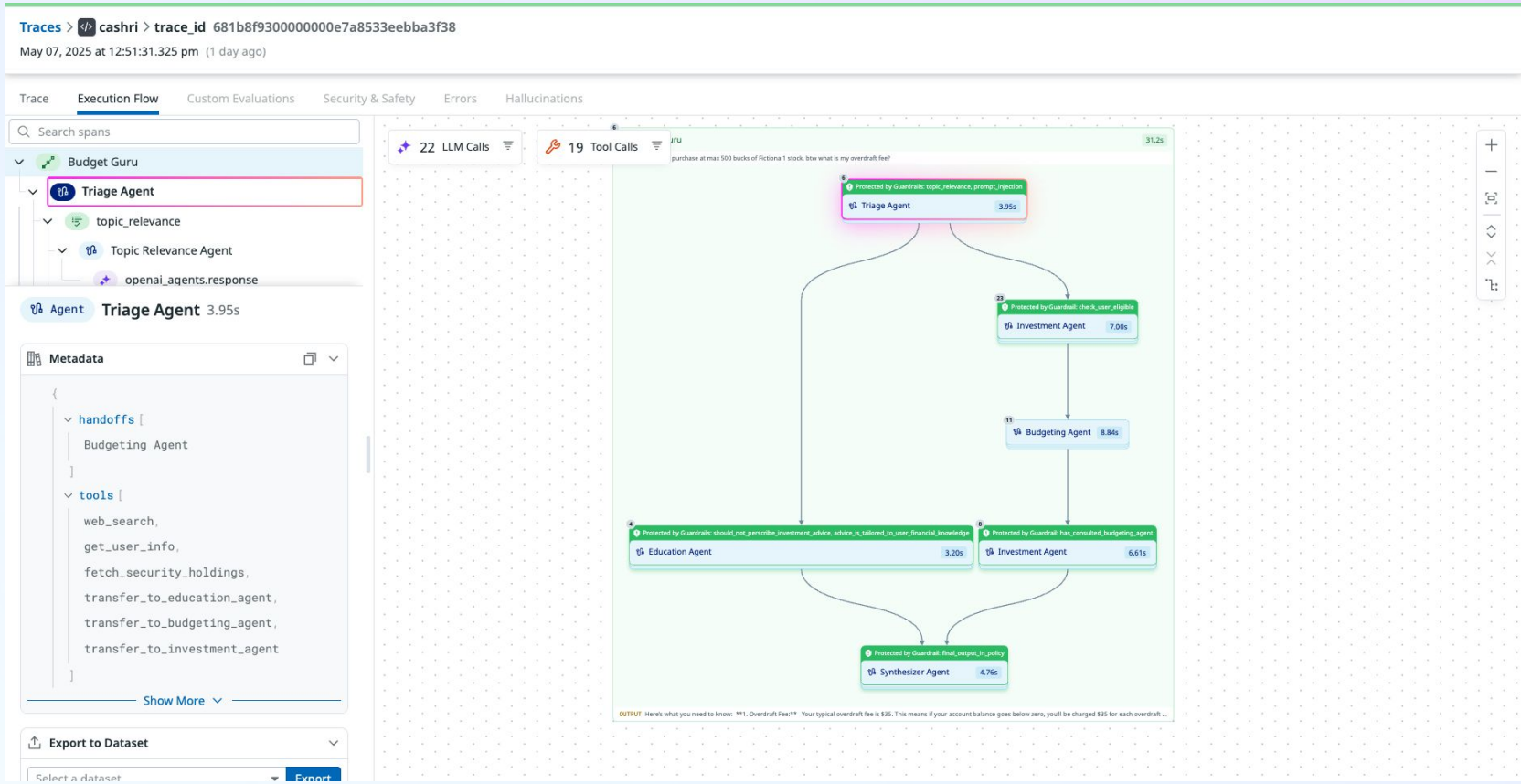
AI Agentic Monitoring

Incluido en LLM Observability

El **Monitoreo de Agentes de IA** permite a los equipos que crean sistemas de agentes — donde agentes toman una variedad de decisiones autónomas y tienen acceso para buscar información — fácilmente analisen y validen sus cargas de trabajo distribuídas. El recurso es actualmente soportado para agentes usando OpenAI Agent SDK, LangGraph, y CrewAI.

Usuarios que lo amarán:

- Ingenieros de IA
- Equipos de SREs / DevOps
- Cientistas de Datos



🎯 La mejor opción para clientes que…

están construyendo aplicativos agentes autónomos de múltiples pasos y se esfuerzan para entender los caminos de decisión y rendimiento de cada paso

👇 Empieza aquí

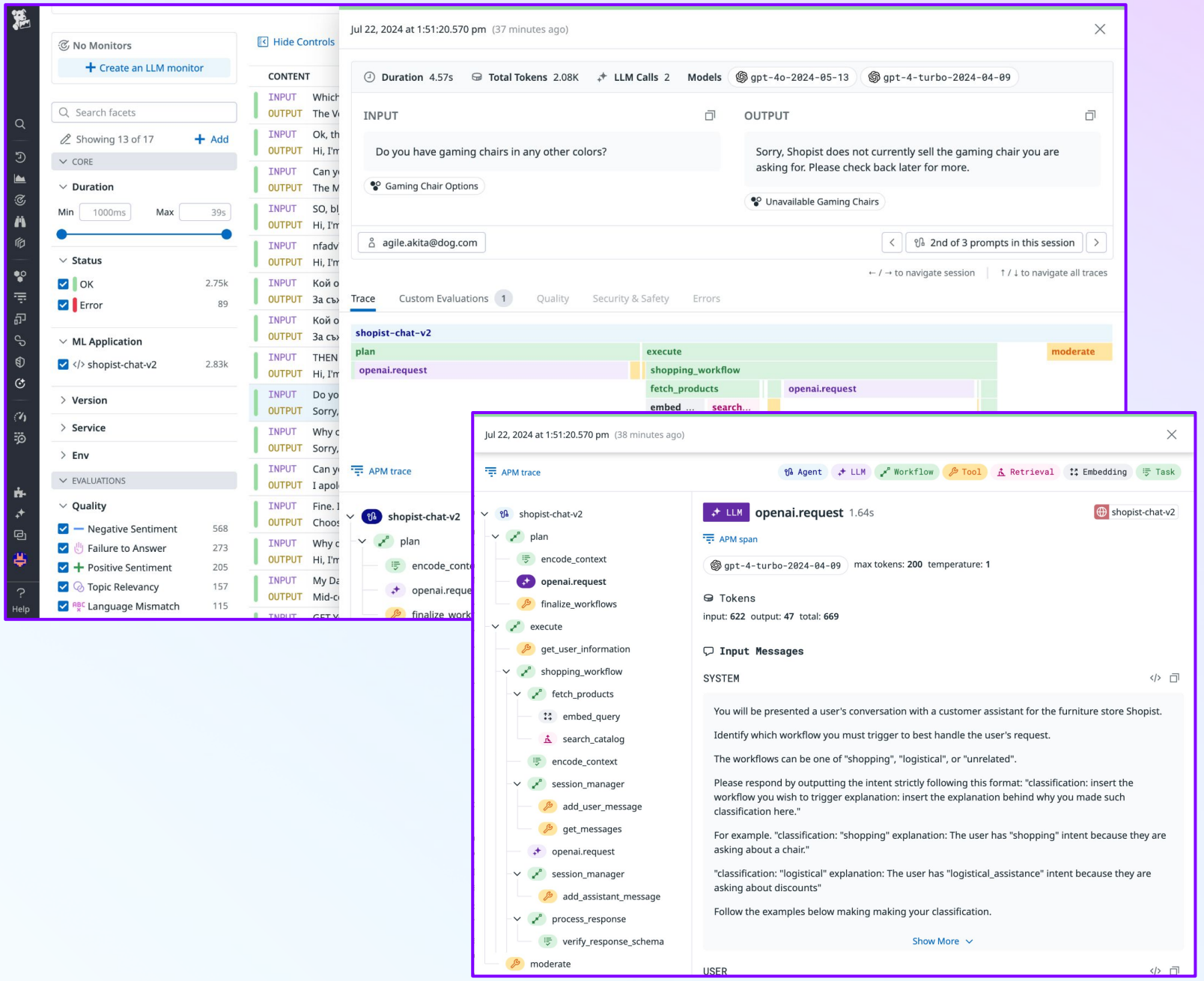
Lee el blog y regístrate para el preview

Mejore el rendimiento de LLMs y reduzca los costos

Detecte errores, fallas y alta latencia en cada paso de la cadena LLM con **traces de extremo a extremo** para cada par de prompt-respuesta

Resuelva problemas como fallas en llamadas LLM, tareas e interacciones de servicio **analizando las entradas y salidas en cada paso** de la cadena LLM

Evalúe la precisión e identifique errores en los pasos de inserción y recuperación para mejorar la calidad y la confianza.

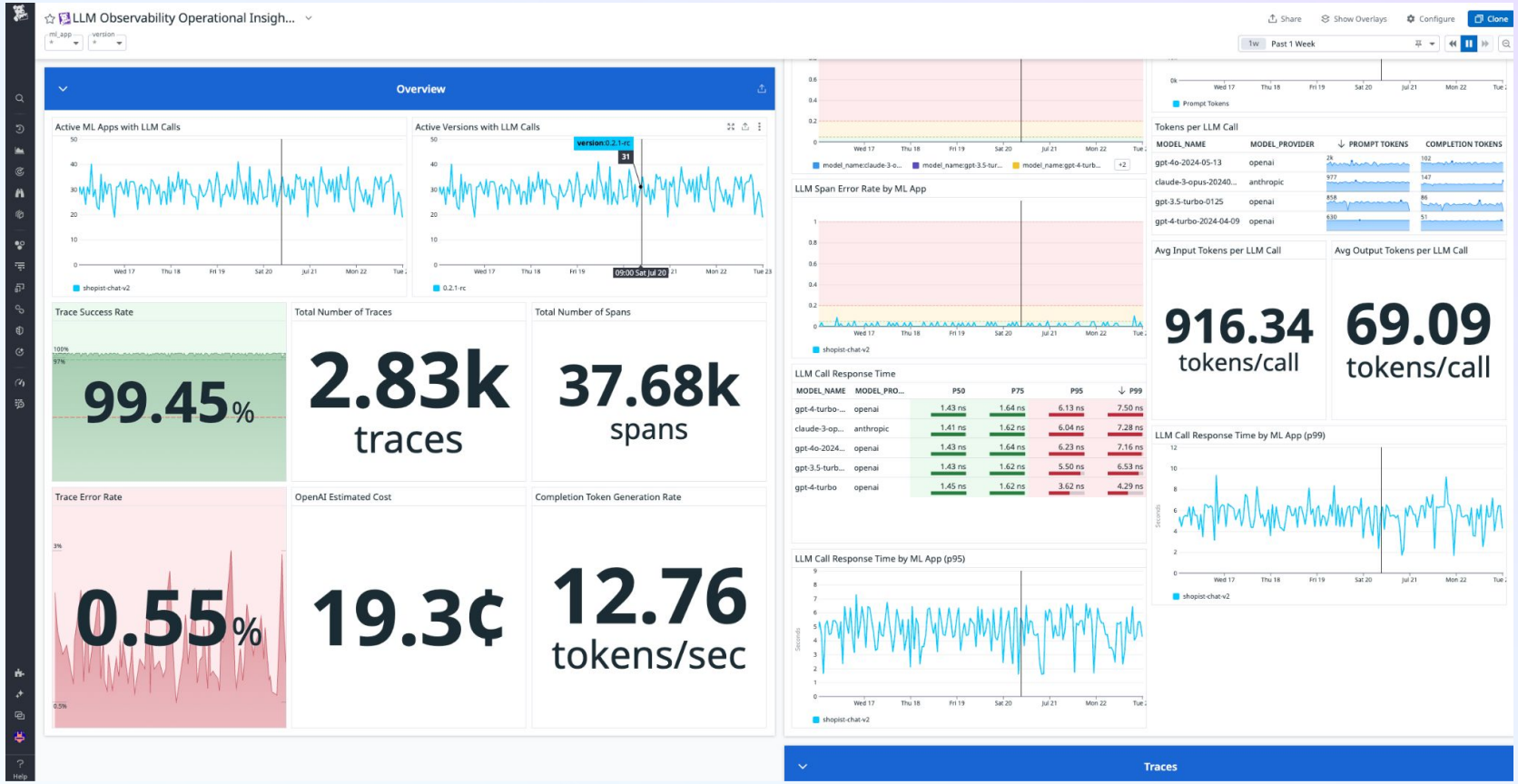


Mejore rendimiento y reduzca costos

Supervise de manera eficiente las métricas operativas clave para las aplicaciones LLM, incluidas las tendencias de costos y latencia, en un **panel unificado**

Descubra instantáneamente oportunidades de optimización de rendimiento y costos con visibilidad de las **métricas de latencia y uso de tokens para cada llamada** en la cadena LLM

Tome medidas rápidamente para mantener un rendimiento óptimo de las aplicaciones LLM con **alertas en tiempo real** sobre anomalías, como picos de latencia o errores



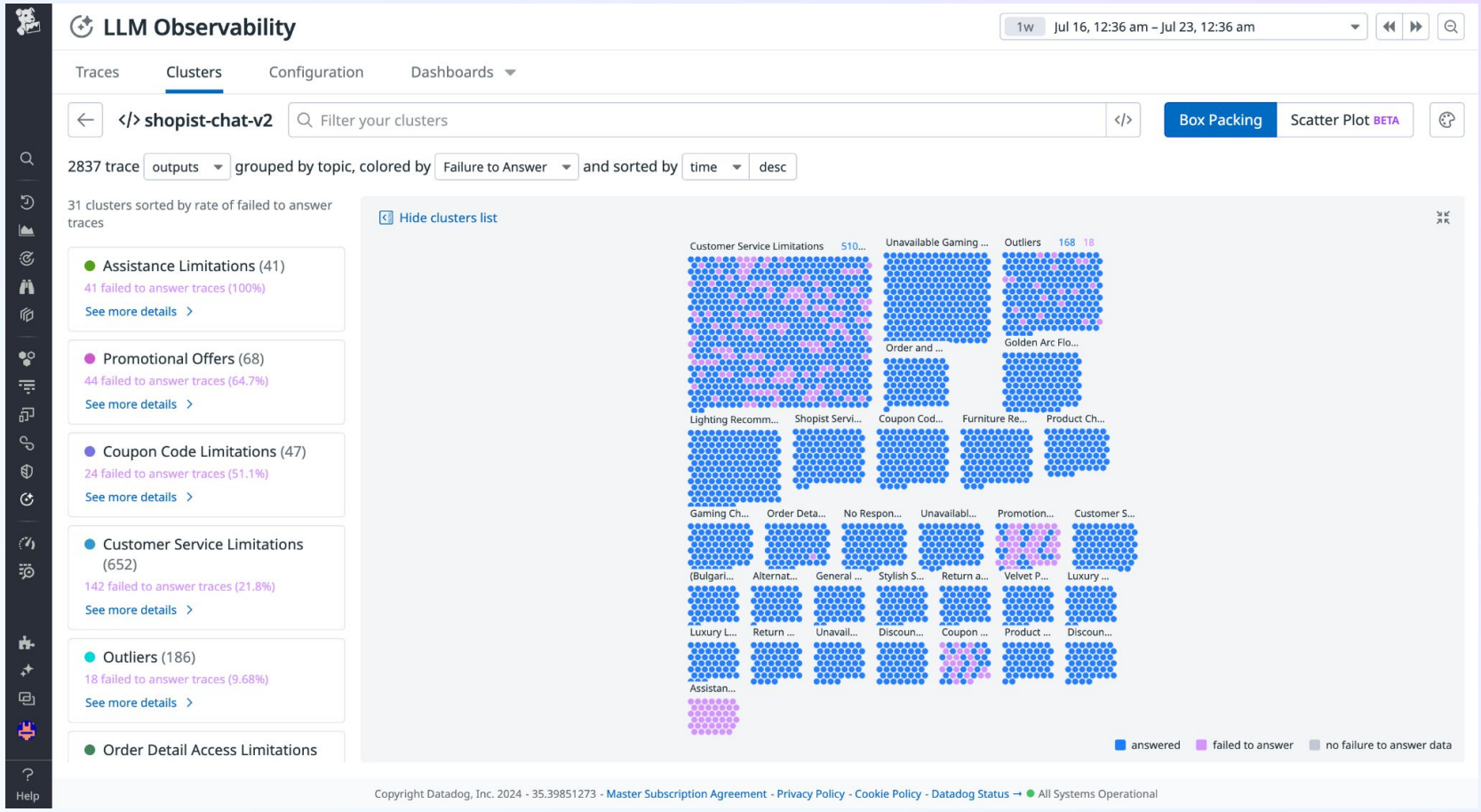
Impulse calidad y seguridad

Mejore la calidad y la seguridad de sus aplicaciones LLM con **evaluaciones personalizadas y preconfiguradas** para detectar inyecciones de prompt, sentimientos de los usuarios, alucinaciones y más.

Optimice los LLM descubriendo desviaciones en la producción mediante el aislamiento de **grupos de prompt-respuesta de baja calidad** semánticamente similares

---

Evite fugas de datos confidenciales, como PII, correos electrónicos y direcciones IP, con **escáneres de seguridad y privacidad integrados** con tecnología de Sensitive Data Scanner.



ANTES

Aplicaciones LLM son una caja negra

Investigación lenta de problemas relacionados a alta tasa de errores, latencia y uso de tokens

Degradaciones de calidad no detectadas y problemas de seguridad

Los cambios degradan el rendimiento, reducen la calidad o aumentan los costos

DESPUÉS

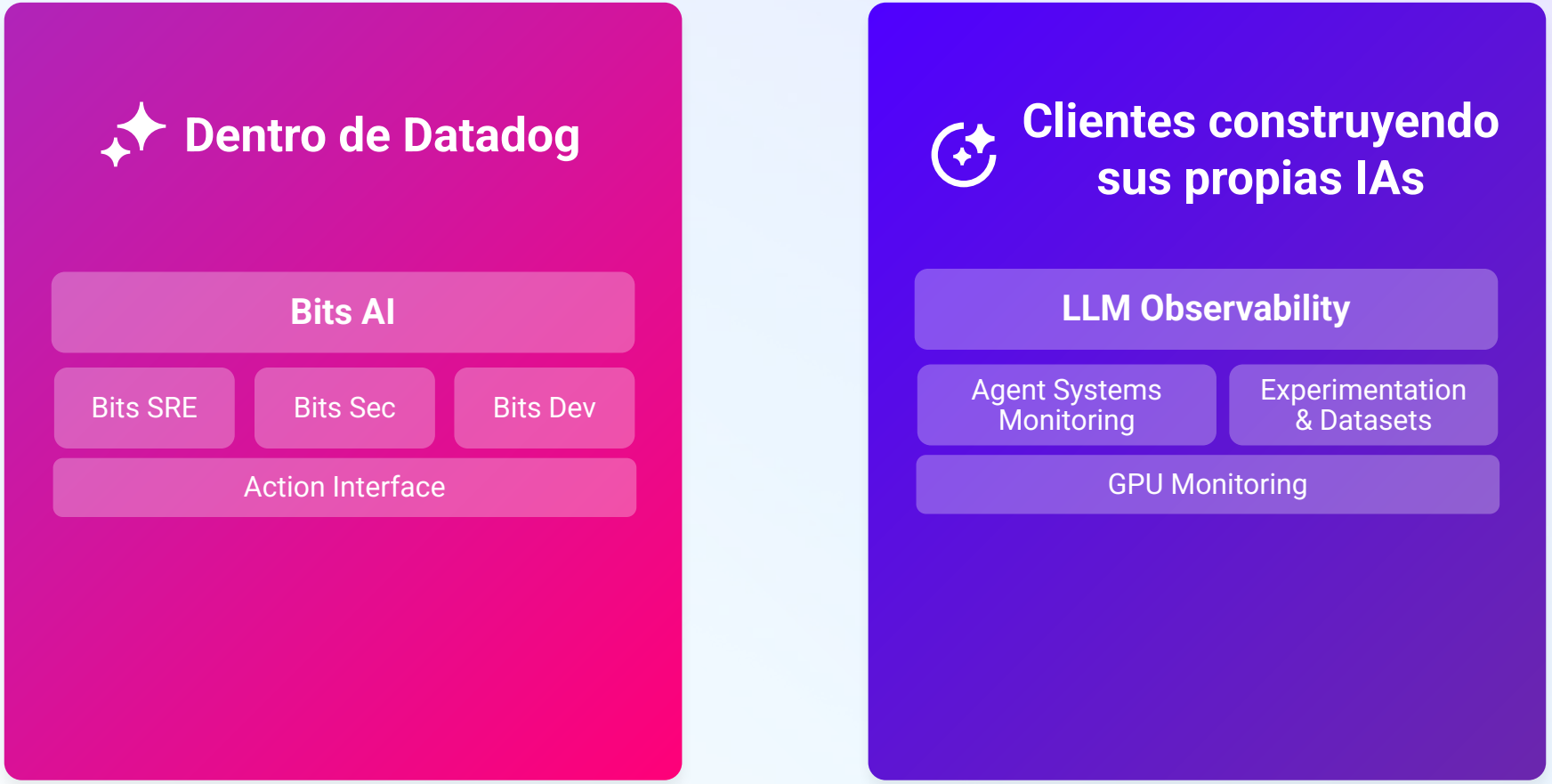
Visibilidad integral de las operaciones de aplicaciones LLM

Rastreo de extremo a extremo para identificar las causas raíz en las cadenas LLM

Evaluaciones personalizadas y OOTB para detectar problemas de calidad y seguridad

Realice experimentos, compare resultados e itere con confianza

Para resumir, estamos hablando de…



Bits AI SRE

¿Para quién es?
SRE, DevOps, Ingenieros de Plataforma

¿Para qué es?
Investigar alertas, crear actualizaciones, coordinar incidentes

Impacto:
Resolver incidentes más rápido, evitar despertar a los ingenieros de guardia, reducir el tiempo medio de reparación (MTTR)

Caso de uso:
¿Alertas críticas a las 3:00 a. m.? Bits AI investiga y sugiere la causa raíz del problema incluso antes de que el ingeniero de guardia abre su laptop.

LLM Observability

¿Para quién es?
SRE, DevOps, ingenieros de IA, cientistas de datos

¿Para qué es?
Aumentar la visibilidad, investigar el rendimiento, la calidad y la seguridad de LLM y controlar los costes

Impacto:
Facilitar la adopción de LLM, optimizar el rendimiento de chatbots y agentes de IA, reducir costes y riesgos de seguridad

Caso de uso:
¿Sus iniciativas de IA están comprometiendo el presupuesto o no cumplen las expectativas de sus usuarios? LLM Observability le ayuda a optimizar sus chatbots y agentes autónomos, manteniendo los riesgos de seguridad y los costes bajo control.

epidata
deploying ideas

DATADOG

Gracias!