NOTA TÉCNICA DETALLADA EN ATENCIÓN AL OFICIO NO. INE/DESPEN/0306/2026

I. Protocolos de Preparación y Verificación de Carga

- ¿Realizaron pruebas de carga (*load testing*) antes del examen? ¿Con cuántos usuarios simultáneos probaron?

  La robustez de la plataforma se fundamenta en cargas probadas previamente, habiendo soportado con éxito aplicaciones con una concurrencia de hasta 80,000 usuarios. Lo que asegura la capacidad operativa necesaria para gestionar el volumen de participantes del examen de poco más de 17,000 usuarios esperados.

  Es preciso puntualizar que el 20 de marzo de 2026 desarrollamos una evaluación con la participación de 14,600 sustentantes y el 4 mayo con 12,660, con las mismas condiciones en las que aplicaría el Instituto Nacional Electoral (INE), lo cual se utilizó como prueba de carga previa al examen.

  Adicionalmente, no se omite señalar que como parte de los procesos de aplicación previo al examen se desarrollan los exámenes de práctica, en este caso los días 08, 12 y 14 Mayo de 2026, funcionando correctamente.

  El análisis técnico posterior permitió identificar que la incidencia no estuvo asociada a insuficiencia general de capacidad, sino al comportamiento específico de los mecanismos de escalamiento automático bajo un patrón particular de saturación de servicios.

- ¿Su equipo de QA (*Quality Assurance* / Control de Calidad) validó el sistema bajo condiciones de estrés?

  Como parte de los procesos de control de cambios CENEVAL, cada vez que se genera una nueva versión del aplicativo se llevan a cabo pruebas unitarias, integrales y de baja carga, que le den validez a la nueva versión.

  Es preciso señalar que a partir de las aplicaciones previamente señaladas no hubo cambios que afectaran la parte arquitectónica para esta aplicación. En los ambientes controlados, no se cuenta con las mismas condiciones de un ambiente productivo (con mismas conexiones de usuarios, tráfico de red, peticiones a los servicios, etc.), por esta razón no se aplican pruebas de estrés en QA.

  En ese sentido, la prueba de estrés se garantizó en aplicaciones previas en productivo, como las de 25 de marzo, 4 mayo, así como los exámenes de práctica, en donde funcionó correctamente la aplicación en cada una de ellas.

- ¿Cuántos usuarios concurrentes esperaban vs. cuántos entraron realmente?

  Se proyectó una demanda máxima de 20,900 sustentantes como fue requerido por el INE en la solicitud de aplicación, registrándose una participación efectiva de 17,291 usuarios durante la jornada; para mayor referencia ver la siguiente imagen.

Dashboard Aplicaciones Semanales 2026 DTIC

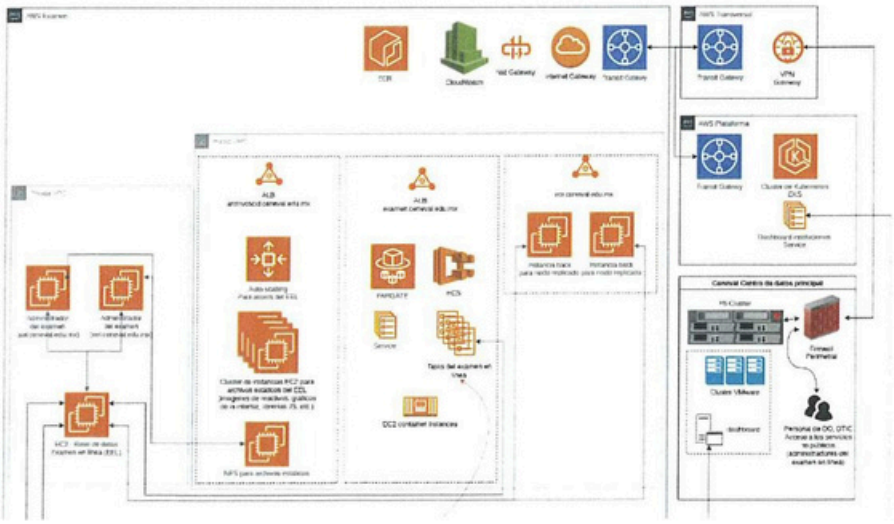
- ¿Tenían un plan de contingencia documentado para sobrecarga del sistema?

  Se contaba con mecanismos operativos y técnicos de contingencia orientados a la continuidad y recuperación del servicio; derivado de la incidencia, dicho protocolo está siendo fortalecido bajo un esquema integral de resiliencia operativa.

II. Infraestructura Tecnológica

- ¿Qué configuración de AWS tienen actualmente? (tipo de servidores, base de datos, capacidad)

  La plataforma opera sobre *Amazon Web Services* (AWS) mediante una arquitectura nativa de nube de alta disponibilidad. Utiliza microservicios en contenedores con orquestación avanzada (ECS/Kubernetes), diseñada para el escalamiento dinámico de recursos según la demanda operativa; para pronta referencia de la estructura se comparte la siguiente imagen:

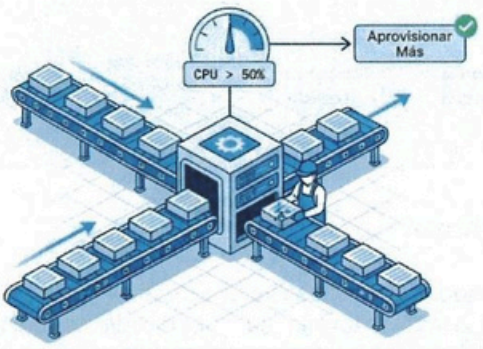
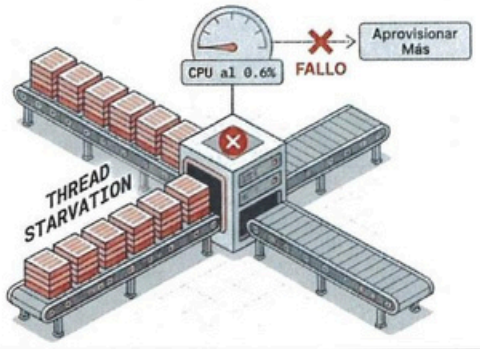


- ¿Tienen auto-escalamiento configurado? Si sí, ¿por qué no se activó o no fue suficiente?

Sí, la plataforma cuenta con auto-escalamiento configurado. Sin embargo, este no se activó automáticamente debido a que las métricas de uso reportadas por los sistemas propios de AWS no alcanzaron umbrales. Durante la aplicación, los valores reportados fueron significativamente inferiores a los consumos reales. Esta lectura impidió que las políticas de auto-escalamiento fueran ejecutadas y, además, evitó que el equipo de monitoreo detectara oportunamente la condición de sobrecarga para una intervención manual.

La política de escalamiento estaba supeditada a un uso de CPU superior al 50%. No obstante, un bloqueo en los hilos de ejecución provocó que el uso de procesamiento descendiera a niveles mínimos (0.6%), invisibilizando la saturación real ante los sistemas de alerta automatizados; para pronta referencia de lo anteriormente explicado, se comparte la siguiente imagen.

La Paradoja del CPU: ¿Por qué no se activó el escalamiento?

El Diseño Teórico (Lo Esperado)	La Realidad Técnica (El Incidente)
	
La política dictaba que si el CPU superaba el 50%, el sistema aprovisionaría más servidores automáticamente.	Los procesos se bloquearon (Thread Starvation), debido a que la política no se activó para el incremento automático de los servidores, esto causó que las tareas se dejaran de procesar y el consumo de CPU cayó drásticamente a 0.6%.

- ¿Cuentan con monitoreo en tiempo real (CloudWatch u otra herramienta)?

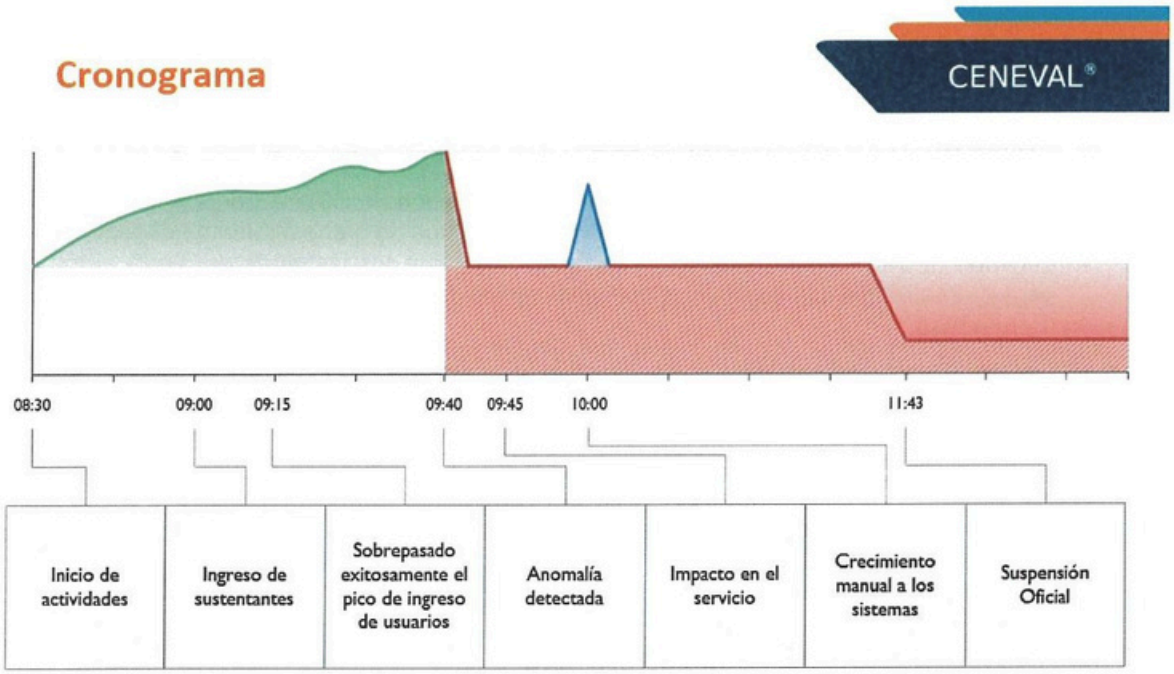
Sí. Se utiliza Amazon CloudWatch en conjunto con tableros de control independientes que supervisan la infraestructura, bases de datos, clústeres de cómputo y sistemas de telecomunicaciones de manera ininterrumpida.

- ¿En qué momento detectaron que el sistema estaba saturado?

Tras un inicio de jornada sin incidencias, a las 09:40 horas se detectó un comportamiento anómalo en los nodos (PODS) encargados de procesar la continuidad del examen. Al no ejecutarse el escalamiento, los recursos se saturaron, derivando en un estrangulamiento de CPU (CPU Throttling) e inanición de hilos, lo que resultó en una degradación progresiva del servicio y latencias severas para el usuario final.

Adicionalmente durante la jornada se recibieron más de 30 mil llamadas, las cuales se dispararon a partir de las 9:40 hrs.

Como parte de la reconstrucción de hechos se comparte la siguiente línea de tiempo:



- ¿Qué acciones tomaron en tiempo real cuando empezó el problema?

Se procedió a una intervención manual directa sobre la infraestructura de servidores. Ante la falla del sistema automatizado, el equipo técnico activó tareas adicionales de escalamiento manual para restablecer la disponibilidad de recursos y estabilizar la operación.

A partir de la identificación preliminar de la incidencia, se inició la revisión integral de la infraestructura tecnológica asociada al Examen en Línea, particularmente respecto de:

- Los parámetros de escalamiento automático.
- La capacidad disponible durante picos de demanda.
- Los umbrales de saturación.
- Los mecanismos de monitoreo en tiempo real.
- Los tiempos de respuesta de servicios dependientes.
- Los protocolos de contingencia ante degradación progresiva del servicio.

Una vez que el servicio ya estaba impactado se trabajó en conjunto con AWS para monitorear el escalamiento del servicio, se restableció, sin embargo derivado de que el sistema se encontraba comprometido, no permitió correr el flujo completo para la aplicación del examen.

III. Análisis de la Incidencia y Salvaguarda de Información

- ¿Por qué los usuarios no podían avanzar entre preguntas? ¿Fue por latencia, timeout, o sobrecarga de la base de datos?

El sistema se encontraba colapsado y se excedió los tiempos de respuesta permitidos, generando errores de comunicación que impidieron la navegación fluida de los sustentantes.

- ¿Por qué no podían volver a entrar? ¿Problema de manejo de sesiones (*session management*)?

  Por la saturación previamente referida.

- ¿La base de datos alcanzó su límite de conexiones concurrentes?

  No. La telemetría confirma que la base de datos mantuvo un desempeño óptimo, utilizando únicamente el 21.92% de CPU y el 60.5% de su capacidad de conexión. El colapso se limitó exclusivamente a la capa de contenedores *Fargate*, es decir la infraestructura para el crecimiento de la base de datos.

- ¿Cuántos usuarios fueron afectados exactamente? ¿Tienen métricas?

  Se registraron para esta aplicación 17,921 e ingresaron al examen 14, 242 sustentantes, de las cuales 825 concluyeron la evaluación y de ese universo 778 respondieron el 100% de las preguntas.

- ¿Se perdió información o respuestas de los usuarios?

  No, la arquitectura del sistema garantiza la retención íntegra de la información. Debido al desacoplamiento entre la capa de cómputo y la de datos, todas las respuestas fueron resguardadas con éxito en la base de datos, asegurando la integridad transaccional del examen pese a las interrupciones en la capa de acceso.