Automatización de la elasticidad del contenedor basada en aplicaciones

Para ingenieros de plataformas y DevOps que buscan operacionalizar la velocidad de comercialización y al mismo tiempo garantizar el rendimiento de las aplicaciones

Índice

Resumen ejecutivo

Su ventaja competitiva depende de la rapidez con que las ideas se conviertan en transacciones empresariales y de lo bien que funcionen para sus clientes. La tecnología es el facilitador.

Los contenedores ofrecen la velocidad, agilidad, elasticidad y escala que está cambiando fundamentalmente la forma en que construimos, implementamos y ejecutamos estas aplicaciones. Marcan el comienzo de un mundo donde las aplicaciones realmente pueden ejecutarse en cualquier lugar; las actualizaciones y nuevas capacidades se pueden implementar en producción varias veces al día, y la demanda dinámica y fluctuante de carga de trabajo se puede gestionar con un suministro de infraestructura elástico, en cualquier lugar y en cualquier momento. Kubernetes es una plataforma que puede permitir que las organizaciones sean ágiles y flexibles, pero no gestiona compensaciones sobre cómo asegurar el rendimiento y al mismo tiempo mantener la eficiencia.

A pesar de toda la simplicidad y agilidad que proporciona la contenerización, la plataforma de orquestación solo proporciona una forma de gestionar el ciclo de vida de estos servicios: implementar y mantener sus servicios de la manera que usted describe.

Las plataformas de contenedores no garantizan de forma nativa que los servicios cumplan con los SLO y no pueden administrar dinámicamente los recursos.
Las políticas basadas en umbrales no resuelven el rendimiento continuo; este método nunca ha funcionado y, dada la velocidad de cambio en las plataformas de contenedores, el escalado automático activado no correlacionado puede terminar causando problemas. La infraestructura elástica es clave para ofrecer rendimiento, pero necesita herramientas de análisis automatizadas que gestionen continuamente la demanda, el suministro y las limitaciones para cumplir los objetivos de nivel de servicio (SLO) deseados.

Este libro blanco analiza conceptos clave que se deben considerar para la adopción de la plataforma de contenedores como la forma de administrar su empresa y para proteger esa inversión con una automatización que asegure el rendimiento mientras minimiza los costes y sigue cumpliendo con las normas. Describe por qué necesita herramientas de análisis de arriba a abajo para que una plataforma Kubernetes autoadministrable ejecute sus servicios. Desarrollar la escala multinube desde el principio de su trayecto le brinda a su organización de TI la “memoria muscular” operativa que transformará fundamentalmente cómo y cuándo entregará más innovación.

Una promesa de velocidad, agilidad, elasticidad y escala.

Kubernetes permite la elasticidad; no garantiza automáticamente que usted cumpla y asegure los SLO de la aplicación.

El éxito en la adopción de la contenerización depende de lo bien que se brinde a los desarrolladores la agilidad que necesitan, la elasticidad necesaria para adaptarse a escala a demandas que fluctúan continuamente y la seguridad de que la aplicación funcionará a la velocidad requerida.

Adoptar un método nativo de la nube y descomponer sus aplicaciones en distintos conjuntos de servicios puede impulsar un desarrollo e implementación de aplicaciones más ágiles. Los contenedores proporcionan el empaquetado que hace que sus servicios sean portátiles y escalables. Kubernetes proporciona un marco y puntos de control para ejecutar sus aplicaciones y servicios digitales. Pero para ofrecer una plataforma eficaz a escala empresarial para su negocio, aún necesita agregar capacidades para liberar la elasticidad de la plataforma que permite cumplir y garantizar los SLO de las aplicaciones.

Implemente más rápido con CICD y feedback de producción.
La metodología adecuada de implementación continua de integración continua (CICD), basada en la automatización, es clave para lograr un tiempo de comercialización más rápido. En el informe de estado de DevOps de Google Cloud 2021,[1] los encuestados citaron mejoras significativas gracias a la implementación de CICD:

Frecuencia de implementación	**Semanal–mensual**	**Cada hora–diaria**
Plazo de entrega del cambio	Más de seis meses	Menos de una hora
Tasa de anomalía del cambio	16%–30%	0%–15%

Con la velocidad surge la necesidad de tener una manera de gestionar el cambio constante en la producción y tener un bucle de feedback sobre el rendimiento de sus servicios y cómo predecir lo que se necesita de la infraestructura. El objetivo es tener una manera de definir sus SLO y hacer que la plataforma brinde feedback sobre cómo configurar sus contenedores e infraestructura para reducir el riesgo de problemas de rendimiento.

- ¿Quién decide cómo se deben asignar los recursos a los servicios? ¿Cómo lo deciden? Pruebas de estrés y benchmarking con SLO establecidos, etc.
- ¿Cómo mide el rendimiento? ¿Existe un bucle de feedback en su ejecución secuencial de CICD para garantizar que los contenedores y los pods estén configurados correctamente?
- ¿Cómo se asegura de que siempre haya suficiente capacidad para nuevas implementaciones?

Encuentre sus respuestas con IBM Turbonomic

Opciones	Limitaciones	Respuestas IBM Turbonomic
Analice manualmente los datos de utilización de contenedores y pods para determinar las especificaciones de recursos.	– Configuración de la recopilación de datos – Mano de obra para el análisis	– Análisis de arriba hacia abajo mediante aplicaciones que determina cómo dimensionar sus contenedores – Comentarios en CICD – Oportunidades para reducir las solicitudes cuando no son necesarias
Analice manualmente los datos de recursos de todos los puntos de la pila para determinar la capacidad de producción.	– Mano de obra para recopilar datos de múltiples orígenes – Mano de obra para el análisis	Análisis basado en la utilización para identificar las necesidades de recursos en toda la pila

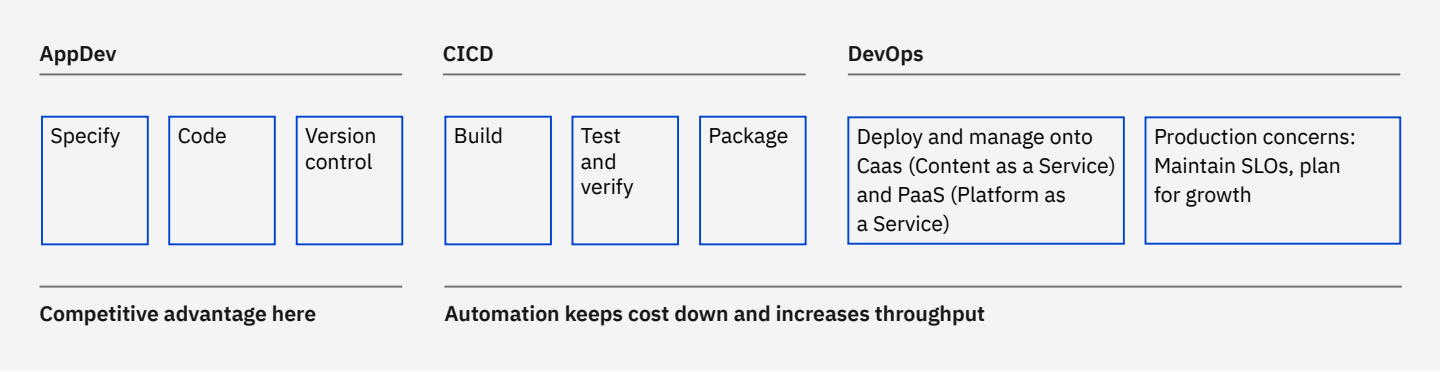


Figura 1. Proceso para agilizar la solicitud.

Plataforma e infraestructura

Por qué necesita una gestión de pila completa basada en aplicaciones
Independientemente de su elección de plataforma de contenedores—o infraestructura subyacente—nube privada, nube pública, nube híbrida, multinube o incluso bare metal—los desafíos operativos de su plataforma como servicio (PaaS) son los mismos:

- ¿Cómo se determina si hay suficiente capacidad para satisfacer la demanda actual y en aumento?
- ¿Cómo se decide cuándo activar más nodos de aplicaciones?
- ¿Cómo se decide cuándo se debe suspender?
- ¿Cómo se maneja la demanda máxima?
- ¿Cómo se utilizan los recursos de la nube pública para la explosión?
- ¿Cómo se garantiza alta disponibilidad (HA) y resiliencia en toda la pila?
- ¿Cómo se hacen cumplir las restricciones comerciales?

La elasticidad habilitada por las plataformas de contenedores brinda la oportunidad de suministrar la suma de las demandas promedio de su aplicación en lugar de la suma de las demandas máximas de sus aplicaciones. Para aprovechar esta capacidad, ofrecer una plataforma que aumente y disminuya continuamente a medida que fluctúa la demanda requiere un software que tome continuamente decisiones de recursos para garantizar que las aplicaciones obtengan el cálculo, el almacenamiento y la red que necesitan cuando lo necesitan.

Encuentre sus respuestas con IBM Turbonomic

Opciones	Limitaciones	Respuestas IBM Turbonomic
Ejecútelo en proveedores de servicios que proporcionen grupos de escalado automático, tales como grupos de servicios afiliados (ASG), conjuntos de disponibilidad, etc.	– Políticas basadas en umbrales – No se puede escalar un nodo específico: todos los nodos deben tener las mismas restricciones, códigos de nodo, etc.	– SLO descendentes basados en aplicaciones – Ajusta continuamente los recursos de infraestructura para satisfacer la demanda de aplicaciones – Escala continuamente hacia arriba, hacia abajo, vertical y horizontalmente, los contenedores, los pods y los nodos correctos – Coloca continuamente los pods en los nodos adecuados
Analice los datos de recursos de todos los puntos de la pila para determinar la capacidad de producción. .	– Mano de obra para recopilar datos de múltiples orígenes – Mano de obra para el análisis	– Análisis basado en la utilización para identificar las necesidades de recursos en toda la pila – Escala continuamente hacia arriba, hacia abajo, vertical y horizontalmente, los contenedores, los pods y los nodos correctos – Activa continuamente acciones para prevenir cuellos de botella

Operando para SLO a escala

El propósito de la plataforma de contenedor es ejecutar sus aplicaciones al nivel de servicio deseado para su negocio. Debe garantizar continuamente el rendimiento a medida que crece el número de aplicaciones. Por lo general, vemos que los clientes tardan más de 12 meses en realizar las primeras 1 o 3 aplicaciones. Para aplicaciones posteriores, con el beneficio de las habilidades adquiridas y las mejores prácticas, puede llevar entre 6 y 12 meses adicionales. Cuando las líneas de negocio aprenden lo que es posible, la escala del número de servicios individuales a gestionar va más allá de la gestión humana. Incluso si ha creado servicios sin estado, aprovechando la naturaleza efímera de los contenedores, ¿cuál es su tolerancia a la degradación del rendimiento para la experiencia del cliente de su usuario final? ¿Qué se puede hacer para gestionar no sólo la demanda, sino también el creciente ritmo de cambio? La respuesta está en la automatización, a través de acciones que se basan en un análisis de las compensaciones de cuántas instancias de su servicio se necesitan para garantizar SLO, la configuración del tamaño y la ubicación de su carga de trabajo y la disponibilidad de recursos compatibles desde la infraestructura.

Los umbrales no resuelven el problema

Una plataforma de contenedor le asegurará tener un número mínimo de servicios disponibles; si uno falla, intentará iniciarlo nuevamente. Pero si desea garantizar una buena experiencia del usuario, querrá que el sistema responda antes de que se degrade el rendimiento y se produzca un fallo. Puede configurar el escalado automático horizontal nativo para satisfacer la demanda, pero debe decidir qué métricas expresan mejor los recursos necesarios, configurar umbrales y límites superior e inferior, probar y extrapolar si funcionará bajo la demanda de producción y luego repetir para cada servicio implementado. ¿Se imagina que tuviera más de 100 servicios para una sola aplicación? Cada una de estas políticas no tiene correlación entre sí. ¿Cómo se puede asegurar de que agregar más pods de un servicio no genera congestión en otra área? ¿Está clonando un pod que estaba mal configurado y necesita escalado vertical primero? ¿Cómo gestiona la congestión de los nodos, tiene en cuenta a los vecinos ruidosos e identifica los recursos asignados no utilizados que podrían liberarse para satisfacer esta demanda?

Además, configurar sus contenedores, pods y políticas de escalado automático de pods horizontales (HPA) o de escalado automático de clústeres no es un ejercicio de una sola vez. Los esfuerzos de estimación deben monitorizarse y redefinirse continuamente si es necesario. ¿Qué podrían hacer sus equipos con el tiempo ahorrado si no tuvieran que configurar y restablecer manualmente estos umbrales?

La importancia de acertar con estas configuraciones tiene implicaciones directas en el lanzamiento exitoso de su estrategia de transformación digital. Algunas malas implementaciones pueden ralentizar significativamente la adopción de las plataformas y sistemas que está creando. Y demasiado tiempo y mano de obra dedicados a configurar manualmente estos puntos de control pueden obstaculizar significativamente la capacidad de su organización para centrarse en la plataforma. ¿Puede su empresa permitirse ese retraso? Lo que se necesita es un sistema de control que pueda gestionar las compensaciones en todos los recursos y definir los límites y las solicitudes de escalado vertical del contenedor, la cantidad de pods necesarios y las decisiones de ubicación para redistribuir los pods y administrar los recursos del clúster utilizando un único motor de análisis.

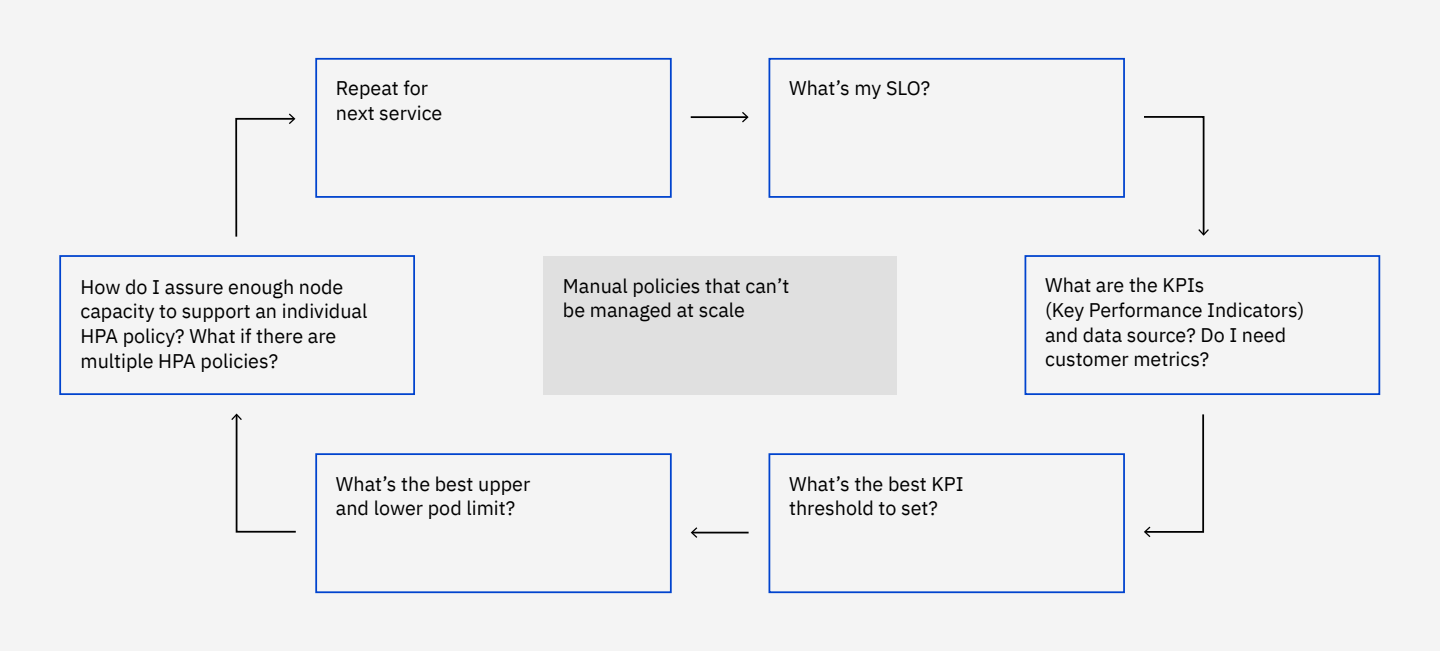


Figura 2. Políticas manuales que no se pueden gestionar a escala

Encuentre sus respuestas con IBM Turbonomic

Opciones	Limitaciones	Respuestas IBM Turbonomic
Política basada en umbrales de HPA sobre cuándo activar los pods de escalado hacia adentro y hacia fuera	– Configura por servicio – Basado en el promedio de todos los pods del servicio – KPI y umbrales definidos manualmente, y límites de pod superior e inferior	– SLO descendentes basados en aplicaciones – Utiliza datos de tiempo de respuesta para impulsar el escalado horizontal de los servicios para cumplir con los SLO – Escala continuamente hacia arriba, hacia abajo, vertical y horizontalmente, los contenedores, los pods y los nodos correctos – Coloca continuamente los pods en los nodos adecuados – Ajusta continuamente los recursos de infraestructura para satisfacer la demanda de aplicaciones
Política basada en umbral de escalador automático de pod vertical (VPA) para escalar verticalmente containers	– Debe definirse para cada servicio – Proyecto Beta: úselo bajo su propia responsabilidad – No accede a la capacidad del nodo para realizar acciones	
Permite que los pods bloqueen para volver a implementarlos en un nodo mejor	Mala experiencia del usuario para transacciones en el pod que está listo para colgarse	
Las soluciones de observabilidad de Prometheus recopilan y consolidan datos	– No proporciona análisis de datos – No proporciona acciones	

Un método basado en aplicaciones

Los SLO de aplicaciones deben impulsar la infraestructura
La contenerización de aplicaciones de misión crítica es una inversión con numerosos beneficios. Pero para aprovechar plenamente esos beneficios de velocidad, elasticidad y portabilidad, necesita un software que le permita tomar las decisiones de recursos correctas en el momento adecuado, 24 horas al día, 7 días a la semana, 365 días al año. De lo contrario, la complejidad le ralentizará.

IBM® Turbonomic Application Resource Management une sus aplicaciones de misión crítica a la plataforma Kubernetes y la infraestructura subyacente, esencialmente dondequiera que se ejecuten sus aplicaciones. Basado en la demanda de las aplicaciones en tiempo real y teniendo en cuenta las restricciones e interdependencias en cada capa de la pila, desde la lógica hasta la física, el software determina las acciones correctas en el momento adecuado para ayudar a garantizar que las aplicaciones siempre obtengan exactamente lo que necesitan realizar. Ejecútelo en tiempo real, planificado o como parte de su ejecución secuencial de DevOps.

Dimensionamiento inteligente: ¿Cómo se deben dimensionar los contenedores?
- Automatice con la implementación: ejecute y mantenga el redimensionamiento como parte de la ejecución secuencial, por ejemplo, YAML, Jenkins, etc.
- Automatice en tiempo real: ejecute dinámicamente a través de Kubernetes.

Colocación continua: ¿Cuándo necesita mover los pods? ¿A qué nodos?
- Ejecute dinámicamente en tiempo real a través de Kubernetes. Solo para servicios sin estado no disruptivos.

Escalado dinámico: ¿cuándo es necesario ampliar o reducir el clúster? ¿En cuanto?
- Ejecute dinámicamente el escalado del clúster en tiempo real a través de infraestructura como código o API de clúster de Kubernetes.

Escalado basado en SLO: ¿Cuándo es necesario ampliar o reducir los pods para cumplir con los SLO de tiempo de respuesta de las aplicaciones? ¿En cuánto?

Requisitos previos para el escalado basado en SLO:
- Las aplicaciones están diseñadas para microservicios horizontales sin estado.
- Tienen una definición y un origen de datos de SLO que Kubernetes no proporciona.

¿Qué significa este tipo de automatización inteligente para usted, sus equipos y su negocio? Los siguientes son los beneficios únicos que ofrece IBM Turbonomic , tanto si está ejecutando Kubernetes en las instalaciones, como en la nube, en Bare Metal Servers o cualquier combinación.

“Control de crucero” para sus aplicaciones: sus equipos establecen SLO de tiempo de respuesta; el software con IA ayuda a garantizar que la plataforma y la infraestructura subyacente siempre proporcionen los recursos que necesitan para cumplir con esos SLO, dondequiera que se ejecuten las aplicaciones.

Minimice el trabajo manual: los desarrolladores, DevOps y los ingenieros de fiabilidad del sitio (SRE) no necesitan establecer umbrales, restricciones o políticas de escalado automático. El software toma las decisiones de recursos correctas por usted y le proporciona acciones que realmente puede automatizar.

No gaste demasiado en capacidad: no es necesario depender de los desarrolladores para tomar decisiones sobre recursos. A menudo suministran en exceso sólo para estar seguros, ¿verdad? Nuestro software determina exactamente qué servicios de recursos se necesitan, todo en función de la demanda de la aplicación.

Acelere DevOps con confianza: aumente de forma segura la frecuencia y la escala de implementación. Nuestro análisis se integra con sus flujos de trabajo de DevOps, lo que contribuye a garantizar que los servicios recién implementados y los existentes siempre funcionen.

Planifique el crecimiento con mayor facilidad: simule la incorporación de nuevos servicios con nuestro software. Determine exactamente cuántos nodos más necesita para respaldar un nuevo crecimiento.

Lo más destacado del cliente

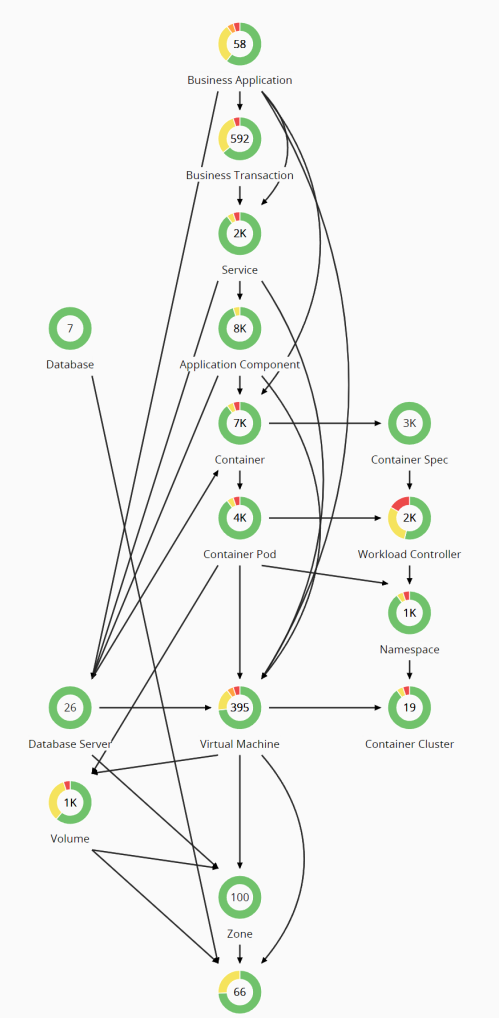
Aceleración de la transformación digital durante una pandemia

Los recursos dinámicos de IBM Turbonomic dentro de la plataforma Kubernetes y la infraestructura subyacente mantuvieron el tiempo de respuesta bajo.

Este cliente es una de las compañías de seguros más grandes de Sudamérica con más de 6 millones de clientes. Su método estándar del sector para gestionar los recursos de los entornos existentes y de próxima generación estaba frenando la transformación digital y la respuesta de la empresa a la pandemia.

La automatización IBM Turbonomic mantuvo el tiempo de respuesta bajo durante el aumento de la demanda durante las fiestas

Este cliente cuenta con una aplicación empresarial que se integra con una de las aerolíneas de bajo costo más grandes que opera en la región. El seguro de viaje se contrata desde esta aplicación, por lo que el pico que vemos en la Figura 3 está relacionado con las vacaciones de Semana Santa de varios días. Si bien la demanda de la aplicación aumentó, los recursos dinámicos de IBM Turbonomic dentro de la plataforma Kubernetes y la infraestructura subyacente pudieron mantener el tiempo de respuesta bajo.



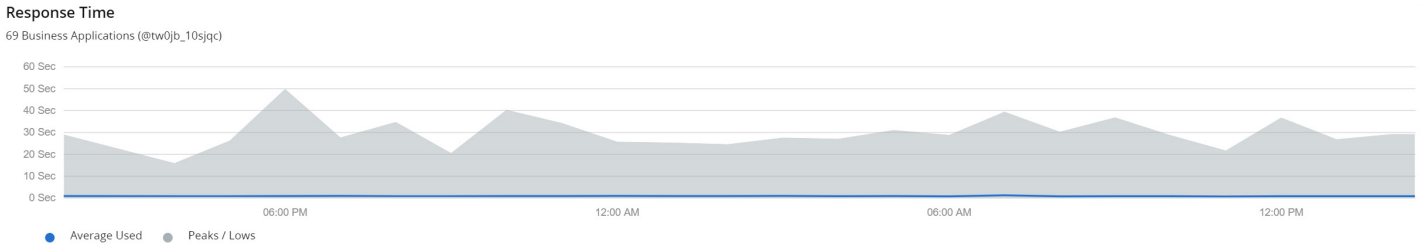


Figura 3. Una vista completa de la aplicación empresarial individual y su tiempo de respuesta con un tiempo de respuesta de automatización mantenido bajo, incluso durante los picos de demanda

57 aplicaciones de misión crítica
- Ejemplo, GPS en coche: denuncia de robo de vehículo, cotización de nuevas pólizas, etc.
- ~3.000 pods (compuestos por ~7.000 contenedores)
- Unido a Dynatrace

Automatizado
- Redimensionamiento del contenedor (transferencia)
- Colocación continua (todos)

~70% de reducción en tíquets

Acerca de IBM Turbonomic, una empresa de IBM

IBM® Turbonomic Application Resource Management proporciona software de gestión de recursos de aplicaciones (ARM) utilizado por los clientes para ayudar a garantizar el rendimiento y la gobernanza de las aplicaciones al proporcionar recursos dinámicos a las aplicaciones en entornos híbridos y multinube. IBM Turbonomic Network Performance Management (NPM) proporciona soluciones modernas Monitoring and Analytics para contribuir a garantizar el rendimiento continuo de la red a escala en redes de múltiples proveedores para empresas, operadores y proveedores de servicios gestionados.

Para obtener más información sobre la automatización inteligente IBM Turbonomic , visite ibm.com/es-es/cloud/turbonomic o hable con un representante de IBM.

IBM España, S.A.
Santa Hortensia, 26-28
28002 Madrid
IBM Corporation
New Orchard Road
Armonk, NY 10504

Producido en los Estados Unidos de América
Marzo de 2022

Este documento se actualizó por última vez en la fecha inicial de publicación e IBM puede modificarlo en cualquier momento. No todas las ofertas están disponibles en todos los países en los que opera IBM.

Los ejemplos de clientes mencionados se presentan únicamente con fines ilustrativos. Los datos reales de rendimiento pueden variar en función de las configuraciones y condiciones de funcionamiento específicas. Es responsabilidad del usuario evaluar y verificar el funcionamiento de cualquier otro producto o programa con los productos y programas de IBM. LA INFORMACIÓN DE ESTE DOCUMENTO SE OFRECE “TAL CUAL ESTÁ” SIN NINGUNA GARANTÍA, NI EXPLÍCITA NI IMPLÍCITA, INCLUIDAS, ENTRE OTRAS, LAS GARANTÍAS DE COMERCIALIZACIÓN, ADECUACIÓN A UN FIN CONCRETO Y CUALQUIER GARANTÍA O CONDICIÓN DE INEXISTENCIA DE INFRACCIÓN. Los productos de IBM están sujetos a garantía según los términos y condiciones de los acuerdos bajo los que se proporcionan.

[1] Estado de DevOps 2021, Google Cloud, 2021