LIBRO BLANCO | JUNIO DE 2016

Más allá del enmascaramiento y la división en subconjuntos: cómo aprovechar el valor de la gestión de los datos de pruebas

Huw Price
CA Technologies

ca technologies

Índice

Sección 1

Introducción

Puede que el concepto de la gestión de los datos de pruebas (TDM) no sea nuevo, pero suele pasarse por alto, infravalorarse y malinterpretarse, mientras que el valor real que puede aportar a la empresa tiende a no aprovecharse. En el caso de muchas organizaciones y distribuidores, se considera únicamente una cuestión del entorno, que equivale a copiar, enmascarar y, posiblemente, dividir en subconjuntos los datos de producción. Estos datos migrados se consideran una "copia final", que se usará en los entornos de desarrollo y control de calidad. En esta coyuntura, el objetivo de una mejor TDM se convierte en la gestión más rápida de esta copia y en satisfacer las necesidades de nuevas copias.

La generación de datos sintéticos —en caso de emplearse— se utiliza de forma aislada y con un enfoque de proyecto a proyecto. Esto suele deberse a la suposición de que la generación de datos constituye una buena práctica para un proyecto o equipo individuales, pero no es viable en toda la empresa. Normalmente, las empresas modernas cuentan con bases de datos enormes y complejas, compuestas por decenas de millones de registros almacenados en numerosos formatos y gestionadas con conjuntos de herramientas dispares. El argumento continúa con la idea de que es necesariamente más sencillo copiar los datos de producción existente que tratar de crear perfiles y modelos de datos tan complejos, como se precisaría para generar datos sintéticos realistas.

No obstante, esta suposición es cuestionable, y las organizaciones que deseen disfrutar de las ventajas de una mejor TDM deben reevaluar la generación de datos sintéticos en el ámbito de toda la empresa, así como el modo empleado para almacenar, gestionar y aprovisionar los datos. La generación de datos sintéticos no solo es más efectiva en cuanto al tiempo, la calidad y el dinero, sino que también —con los cambios de tecnología, procesos y estructuras de equipos adecuados— tiende a constituir una estrategia más sencilla y segura que el enmascaramiento completo de los datos de producción.

En este documento se contrastan los problemas habituales con las posibles ventajas de la TDM desde la perspectiva del entorno; además, se consideran los procesos, la tecnología y las estructuras de equipos que suelen darse en las organizaciones[1].

Sección 2

Las desventajas de un enfoque puramente logístico para TDM

En las organizaciones que utilizan datos de producción para entornos de desarrollo y pruebas, el enmascaramiento y la división en subconjuntos representan prácticas obligatorias. El enmascaramiento resulta necesario para cumplir la legislación actual sobre la protección de los datos, mientras que la división de los datos de producción en subconjuntos puede contribuir a reducir unos costes de infraestructura inviablemente elevados.

Sin embargo, si las prácticas de TDM comienzan y terminan por el enmascaramiento y la división en subconjuntos, resultará imposible materializar el valor real que puede aportar a una empresa. Esto se puede demostrar de acuerdo con dos suposiciones: primero, que los datos recopilados por cualquier organización moderna contendrán información de identificación personal que, por ley, no puede abandonar los entornos de producción de forma reconocible; segundo, que los datos enmascarados están destinados a fines de desarrollo y pruebas y, por lo tanto, deben abarcar todos los casos de prueba que se deban ejecutar, así como cualquier funcionalidad nueva.

Tras estas suposiciones, primero se demostrará por qué el enmascaramiento no es en realidad ni la forma más sencilla ni la más eficaz de aprovisionar datos seguros idóneos para los entornos de desarrollo y pruebas. A continuación, se analizarán las estructuras de equipos, la tecnología y los procesos implementados en la mayoría de las organizaciones y se aducirá que, si solo se utilizan el enmascaramiento y la división en subconjuntos, la cobertura deficiente de los datos de producción —junto con la forma en la que se suelen gestionar— supondrá la persistencia de los puntos débiles habituales (retrasos en los proyectos, excesos presupuestarios y defectos en la producción).

El enmascaramiento no es la solución sencilla

En primer lugar, es posible discutir la idea de que la copia y el enmascaramiento seguros de los datos de producción para entornos de desarrollo y pruebas constituye la opción más simple en una infraestructura de TI compleja. En realidad, el esfuerzo de enmascarar los datos de producción a la vez que se conserva su integridad referencial tiende a superar el de la creación de modelos de los datos. Y, de todas formas, para llevar a cabo un enmascaramiento efectivo, primero se deben haber generado perfiles de los datos. Además, resulta habitual no enmascarar determinados aspectos complejos de los datos como una especie de compromiso, por lo que el enmascaramiento podría no constituir tampoco la opción más segura en muchos casos.

A la hora de enmascarar los datos para su uso en entornos que no sean de producción, se deben mantener las relaciones entre las columnas de los datos originales, incluso cuando se enmascara el contenido confidencial. En un nivel simple, esto quiere decir que se conserva la integridad referencial. En cambio, en un nivel más complejo, significa que se mantienen las relaciones complejas entre columnas. Por ejemplo, si existe un campo "Total" que calcula el total de dos o más columnas, estas últimas tendrán que mantenerse conectadas en los datos enmascarados. Las relaciones más complejas son las de datos temporales y las causales —por ejemplo, un total basado en tiempo—, cuyo enmascaramiento homogéneo entraña una dificultad extrema.

Es más, estas relaciones no solo se deben enmascarar de forma homogénea dentro de una base de datos (intrasistema), sino que, en una organización de gran tamaño típica con varios tipos de bases de datos y aplicaciones complejas y compuestas, también tendrán que mantenerse entre las distintas bases de datos (intersistema). Cuanto más complicados sean los datos, más difícil resulta, mientras que, al mismo tiempo, los datos pasan a ser más sencillos de descifrar y hay más información que debe correlacionarse.

En la mayoría de los casos, el enmascaramiento de datos se centra principalmente en el contenido y en las relaciones intrasistema e intersistema, en lo que respecta a conservar la integridad referencial. La complejidad de mantener el contenido, las relaciones de columnas y las relaciones temporales al mismo tiempo tiende a traducirse en que uno de esos aspectos se descarte. Por ejemplo, con frecuencia las relaciones temporales entre las columnas no se enmascararán. Aunque, desde el punto de vista legal o técnico, no se trata de contenido confidencial, estos registros preservados pueden utilizarse para identificar información confidencial correlacionándola con datos ajenos a ella.

Pongamos por ejemplo un conjunto de registros de transacciones enmascarados y un atacante que sabe que una persona concreta ha efectuado una transacción con un importe determinado en una hora específica. Aunque esta información no se encuentra en el contenido confidencial, que se ha enmascarado, la información temporal podría seguir estando presente en la base de datos enmascarada, ya que el enmascaramiento homogéneo de contenido, totales numéricos y horas de las transacciones entraña una dificultad extrema. Una vez que se haya identificado la información confidencial en la primera instancia, se podrá hacer lo propio en todos los sistemas y las bases de datos, ya que se ha conservado la integridad intersistema e intrasistema. La consecuencia de todo esto es que los datos dejarán de presentar un carácter anónimo.

Aunque se haya realizado el esfuerzo de generar perfiles de los datos y enmascararlos, con toda su complejidad, estos aún no se podrán considerar adecuadamente protegidos, ya que se seguirá conservando información en sus aspectos más complejos. Por ejemplo, un atacante podría identificar los efectos de cascada causales de un registro de transacciones combinando la información causal, de intersistema y de intrasistema con métodos temporales, como una comparación de instantánea. Por tanto, sería posible deducir cómo se han cambiado los datos durante una operación, así como cualquier efecto de cascada causal que puedan estar presentes (p. ej., cualquier desencadenador que se active en la base de datos). Reiteramos, la seguridad de los datos enmascarados equivale únicamente a la de su eslabón más débil: una vez que se haya identificado esta información, un atacante puede valerse de las distintas relaciones entre columnas para descifrar requisitos confidenciales desde el punto de vista comercial o, peor aún, información de identificación personal.

Si se tiene en cuenta la suposición de que prácticamente todos los datos de una organización contendrán alguna información de identificación personal, el enmascaramiento no presenta un modo seguro ni eficaz de aprovisionar los datos a los entornos que no sean de producción. Se trata de una conclusión especialmente válida, dadas las crecientes exigencias que afrontan las organizaciones para proteger sus datos por completo: la multa media por una infracción de datos aumentó en un 13 % para situarse en los 3,5 millones de dólares en 2013 [5, Ponemon, 2014]. Por su parte, la próxima normativa de protección de datos general de la UE se perfila como un mecanismo que permitirá una mayor aplicación de la legislación existente que prohíbe el uso de la información personal para cualquier finalidad que no sea por la que se recopiló.

Si también incluimos en esta ecuación que el informe sobre el estado de la privacidad de 2015 (State of Privacy Report 2015) de Symantec descubrió que las preocupaciones por la seguridad de los datos determinan dónde y cómo el 88 % de los consumidores europeos realizan sus compras, el riesgo de usar datos enmascarados en entornos que no sean de producción resulta inviable. Ahora que se han establecido los riesgos legales de la migración de los datos a entornos de desarrollo y pruebas, podemos plantearnos su eficacia y eficiencia.

Personas

Dependencias de datos

No es habitual que una organización cuente con un equipo centralizado a cargo de la TDM; en lugar de ello, los equipos individuales tienen la responsabilidad de gestionar, encontrar y enmascarar sus propios datos. Por lo tanto, los equipos trabajan aislados los unos de los otros, en sus propios entornos de desarrollo o pruebas, pero tienden a utilizar las mismas fuentes de datos. Esta falta de centralización y colaboración conlleva restricciones de dependencias de datos, según las cuales, si un equipo realiza un cambio en una base de datos, todos los equipos se ven afectados. Esto se traduce en frustraciones, retrasos y reprocesamientos, ya que las pruebas fallan por ningún motivo aparente y los equipos no pueden identificar si se debe a un defecto del código a un error en los datos. Los equipos también tienden a carecer de la capacidad de versionar o parametrizar los datos de pruebas, al igual que la de revertir las bases de datos si otro equipo las destroza.

El SDLC se continúa observando secuencialmente, como una serie de fases lineales: se avanza a la siguiente fase cuando un equipo finaliza su tarea. En consecuencia, suele ser habitual que los equipos de desarrollo y pruebas vean sus proyectos paralizados debido a retrasos en etapas previas del proceso. Deben esperar a que los datos estén disponibles o a que otros equipos completen las especificaciones y fuentes de datos; por ello, una cantidad considerable del tiempo del SDLC puede dedicarse a esperar la recepción de datos.

Esta falta de paralelismo contrasta fuertemente con las iniciativas actuales en muchas organizaciones para obtener una entrega continua, en las que los equipos necesitan datos en todas las etapas del SDLC. Por ejemplo, esta coyuntura hace que resulte imposible llevar a cabo el desarrollo de fin de vida en versiones antiguas de un sistema, al mismo tiempo que también se utilizan datos de las versiones nuevas. En su lugar, los equipos se encuentran con un nuevo entorno de desarrollo, el cual carece de los datos necesarios.

Tecnología

Dependencias del sistema

Además de las dependencias de datos, las restricciones del sistema y hardware representan un punto débil habitual para los equipos de desarrollo y pruebas modernos. Durante las últimas dos décadas, las aplicaciones han ido aumentando en complejidad, ya que presentan una variedad de dependencias en continua expansión con otros sistemas, tanto dentro como fuera de la organización. Esta situación suele conllevar obstáculos cuando se desea probar un sistema, ya que los servicios podrían ser inestables o no encontrarse disponibles. Es posible que los equipos descubran que otro equipo goza de prioridad en el entorno que desean usar o que no puedan obtener los datos correctos, ya que, según los equipos de desarrollo y pruebas, existe una falta de entornos de tamaño completo. Aunque al principio pueda parecer que estas limitaciones carecen de una relevancia directa para los datos, más adelante se explicará cómo una mejor infraestructura de TDM constituye un paso necesario para su resolución.

Almacenamiento de datos

Las organizaciones modernas almacenan un gran volumen de datos de producción. Tener copias de estas bases de datos y ejecutarlas en los equipos de desarrollo resulta caro y lento. El almacenamiento de tales datos conlleva un enorme coste de infraestructura, entre los que se cuenta el coste del hardware, las licencias y el soporte. Además, los servidores se ven sometidos a exigencias cada vez mayores, ya que deben transmitir altos volúmenes de datos a diversas tareas simultáneamente, a la vez que ejecutan conexiones abiertas, mantienen archivos abiertos y gestionan inyecciones de datos[2].

A menos que se reevalúe la forma en la que se gestionan los datos, es improbable que estos costes disminuyan. La cantidad de datos que recopila y almacena una empresa media se duplica cada año[3], y la llegada de los "grandes datos" ha conllevado que en las organizaciones ahora se hable de gestionar petabytes, no terabytes. En consecuencia, el almacenamiento de datos se ha convertido en una de las partidas de los presupuestos de TI que más crecen, y existen indicios de que el sector del almacenamiento de datos ha disfrutado hace poco de un crecimiento anual de nada menos que un 20 %[4]. Por lo tanto, las organizaciones deben considerar si cada copia de los datos de producción resulta necesaria; dado que algunas organizaciones tienden a contar con varias copias de una sola base de datos, es probable que la respuesta sea "no".

Creación de perfiles y extracción de datos

Sin una tecnología que posibilite su automatización parcial o completa, la detección de datos representa uno de los mayores retos que afrontan los equipos que desean ofrecer software totalmente probado a tiempo, especialmente en un contexto de desarrollo ágil o en un marco de trabajo de entrega continua. Los evaluadores pueden dedicar la mitad de su tiempo a buscar datos, lo que les obliga a encontrar un compromiso no deseado entre la ejecución de todas las pruebas necesarias para evitar que los defectos costosos transpiren a la fase de producción y entregar el software a tiempo.

El almacenamiento heterogéneo de datos en hojas de cálculo no controladas (con referencias cruzadas e indexado escasos o nulos, por ejemplo), en lugar de en un almacén o repositorio centralizado, no hace más que empeorar la situación. Además, las bases de datos no suelen estar adecuadamente documentadas, ya que las organizaciones tienden a carecer de diccionarios centrales de los atributos de los datos de pruebas y las consultas de SQL relacionadas. Por lo tanto, la extracción y asignación de datos se ven perjudicadas, y los equipos no pueden solicitar datos mediante plantillas o formularios estandarizados ni según los criterios específicos que necesitan. En consecuencia, con frecuencia tendrán que encontrar manualmente un pequeño conjunto de datos aptos que se ajuste a los requisitos y casos de prueba individuales, lo que constituye un proceso arduo y propenso a errores.

Asimismo, la falta de herramientas automatizadas de extracción y creación de perfiles de datos aumenta el riesgo de la no conformidad. Como los datos se almacenan en hojas de cálculo no controladas, puede haber información confidencial en cualquier lugar: por ejemplo, en una columna de notas, cuando no hubiera un campo de columna adecuado para los datos o si todos los campos relevantes ya estaban rellenos. Sin poder buscar campos específicos, existen menos probabilidades de que se encuentre dicha información y, por lo tanto, cabe la posibilidad de que no llegue a los entornos que no sean de producción. Esto va en contra de la normativa de protección de datos, donde se especifica que los datos solo se pueden usar para la finalidad con la que se recopilaron. La empresa corre el riesgo de recibir cuantiosas multas que promedian 3,5 millones de dólares[5], lo cual puede perjudicar gravemente su balance final.

Proceso

Los datos de producción no constituyen una "copia final"

La incapacidad de encontrar datos adecuados para los requisitos y casos de prueba nos conduce al mayor problema derivado del uso de datos de producción en entornos que no son de producción: sencillamente, no pueden cumplir la mayoría de las solicitudes de datos que una organización suele recibir, en especial las que provienen de los entornos de desarrollo. Como ya hemos descrito, los equipos de desarrollo y pruebas necesitan datos en todas las etapas del desarrollo. Si lo consideramos adecuadamente, una base de datos de "copia final" debe contener un conjunto estándar de datos con

los que realizar pruebas de forma reiterada, así como los datos necesarios para satisfacer todas las evaluaciones posibles. Además, deben ser similares a los de producción y estar actualizados, e incluir "datos incorrectos", así como toda la información previa. Los datos de producción solo cumplen dos de estas condiciones: son similares a los de producción y podrían contener toda la información previa. Por lo tanto, no se trata de una "copia final".

Una gran parte de los datos de producción son muy similares: abarcan transacciones de "negocio habitual" y, por su propia naturaleza, excluyen los datos que provocarán que el sistema se colapse. Por lo tanto, no contienen datos incorrectos, lo que conllevará que no se evalúen rutas negativas y nuevos escenarios durante las pruebas. Sin embargo, estos resultados inesperados, singularidades y condiciones límite son precisamente los que habitualmente ocasionan que un sistema se colapse, y el objetivo del desarrollo y las pruebas debe ser comprobar estos casos extremos, rutas inadecuadas y escenarios inesperados. Al utilizar únicamente métodos de muestreo, transpirarán defectos a la fase de producción, donde su corrección puede costar hasta 1000 veces más[6] y su resolución puede llevar 50 veces más tiempo[7].

Asimismo, los datos muestreados de la producción casi nunca estarán actualizados, dados los cambios del entorno y las exigencias empresariales en constante cambio que afrontan los equipos de TI. Como los datos no se almacenan independientemente de los entornos, cuando uno de ellos cambia, se precisa una actualización del sistema o de la versión. Esto puede destrozar los datos y escenarios que se han creado a partir de varias fuentes de datos de producción, y se pueden perder conjuntos de datos útiles, lo cual se tendrá que recrear laboriosamente a mano. Además, estas actualizaciones del sistema tienden a ser sumamente lentas. Por ejemplo, hemos trabajado con organizaciones donde hemos observado que han invertido hasta 9 meses en completar este proceso.

La creación de datos manualmente suponer una solución a corto plazo, que permite que se comprueben los casos de prueba inmediatos. Sin embargo, como dichos datos se generan de acuerdo con requisitos y casos de prueba específicos, se quedarán obsoletos casi de inmediato. Podemos encontrar un buen ejemplo de esto en los tipos de cambio de divisas o en las tendencias bursátiles. En estos casos, los datos se quedan obsoletos a diario; es decir, los datos creados manualmente no se pueden volver a usar, de modo que se suelen descartar. Después, se deben crear arduamente nuevos datos para cada prueba, lo que conlleva aún más retrasos en los proyectos, mientras que las pruebas se posponen hasta la siguiente iteración o se transfieren a la siguiente fase del ciclo de vida de desarrollo.

Si para un enmascaramiento efectivo en varias bases de datos se precisa que primero se creen perfiles de los datos —y aun así no se garantiza la seguridad ni se proporcionan datos de la suficiente calidad—, debemos plantearnos la siguiente pregunta: ¿por qué, una vez que se han creado perfiles de los datos de producción correctamente, las organizaciones no optan por sencillamente generar sintéticamente los datos necesarios?

Sección 3

La alternativa ideal: personas, procesos y tecnología

Para disfrutar de una política de TDM mejor, se necesita un enfoque estructurado y, especialmente, centralizado para gestionar los datos de toda la empresa. Con este planteamiento, no solo se solucionan muchos de los puntos débiles indicados con anterioridad, sino que, en realidad, suele constituir un modo más barato, eficiente y sencillo de aprovisionar datos a los entornos de desarrollo y pruebas —siempre y cuando se cuente con la tecnología adecuada—.

Al almacenar datos útiles como activos fuera de los entornos y transferirlos a estos últimos a petición, se solucionan los problemas de entorno de la TDM, que pasa a preocuparse por cómo se devuelven los datos. Esto, a su vez, puede realizarse de forma eficiente con un enfoque centralizado para el almacenamiento de datos, según el cual, los datos se modelan como activos reutilizables, que se pueden extraer a petición como subconjuntos precisos. En lugar de contentarnos con

el enmascaramiento y la división en subconjuntos como operaciones necesarias, la generación de datos sintéticos presenta un objetivo estratégico que, si se incorpora en una política de TDM más amplia, posibilita la entrega de software completamente probado, a tiempo y sin superar el presupuesto.

Tecnología

Creación automatizada de perfiles de datos

Ya se ha sostenido que, para enmascarar los datos de producción de forma eficiente, primero se deben crear perfiles de ellos; pero incluso entonces, los datos enmascarados no serán completamente seguros ni ofrecerán la información necesaria para el desarrollo y las pruebas. No obstante, existen tecnologías automatizadas que reducen el esfuerzo necesario para generar perfiles de los datos en un panorama de TI complejo. Para crear perfiles de los datos, primero hay que "registrarlos"; es decir, recopilar tantos metadatos como resulte posible. Entre tales metadatos se incluyen nombres de tablas y de columnas, y tipos de columnas, por ejemplo. Además, estos existen incluso en sistemas de bases de datos no relacionales, en forma de ejemplares de sistemas de mainframe y documentos de asignación para archivos sin formato delimitados o de anchura fija.

Una vez que se haya efectuado el registro, se pueden aplicar algoritmos matemáticos de detección de datos. Con CA Test Data Manager (antes, Grid-Tools Data Maker) por ejemplo, esto se realiza primero para los esquemas individuales; se determina la información de identificación personal y se aplican técnicas de ingeniería inversa a las relaciones de la base de datos, si resulta necesario. Una vez que se ha efectuado este paso, es posible unir los sistemas. De este modo, CA Test Data Manager utiliza vistas de cubos para generar perfiles de incluso las relaciones más complejas que existen en los datos, con lo que crea un conjunto de datos multidimensional en el que cada dimensión del "cubo" representa un atributo de los datos. Esta generación de perfiles permite a las organizaciones saber exactamente qué datos existen y dónde están almacenados, así como identificar cualquier laguna en su cobertura funcional.

Generación datos sintéticos

Tras haber creado una perspectiva precisa de los datos que existen e identificado qué información adicional se necesita para los entornos de pruebas, es posible generar automáticamente cualquier dato que falte. Como cada mundo real puede considerarse otro punto de datos, es posible modelar y crear datos para cubrir la totalidad de las variaciones funcionales. Estos datos incluyen escenarios futuros que no han ocurrido antes, así como "datos incorrectos", singularidades y resultados inesperados. Así, se logran unas pruebas negativas efectivas y el desarrollo de un nuevo sistema o subsistema. Proporciona una forma sistemática de efectuar las pruebas, de modo que los resultados inesperados y escenarios que podrían no ocurrírseles a los evaluadores no provoquen el colapso del sistema; además, a la vez, se detectan los defectos antes de que transpiren a la fase de producción.

Si no hay suficientes datos con los que efectuar pruebas reiteradas, también se pueden crear grandes volúmenes de datos mediante una tecnología automatizada. CA Test Data Manager proporciona una herramienta automatizada que funciona directamente con capas de API de ERP o RDBM para generar datos tan rápido como lo permita la potencia de procesamiento. Sus scripts de procesos masivos pueden duplicar la cantidad de datos que posee una organización con la misma rapidez que permita la infraestructura. En resumen, esta tecnología automatizada permite la rápida creación de datos parecidos a los de producción, con el volumen necesario para ejecutar las pruebas, incluidas las negativas, y datos suficientes para realizar pruebas reiteradas.

Gestión de los datos centralizada

Con las mejoras tecnológicas mencionadas, una organización ya habrá recorrido una gran parte del camino para establecer una "copia final", tal y como se define en el presente documento (consulte la sección Los datos de producción no constituyen una "copia final"). Al almacenar los datos, modelados como objetos reutilizables, en un almacén de datos de pruebas central, también podrán disfrutar de la capacidad de identificar rápidamente subconjuntos específicos de los datos para los entornos de desarrollo y pruebas, a petición.

Clonación de datos

Una vez que se hayan modelado los datos como objetos en un "mercado de datos" o almacén de datos de pruebas, y se hayan establecido los diccionarios de activos de datos y consultas asociadas, será posible identificar, clonar y enviar a entornos de desarrollo y pruebas subconjuntos de datos específicos.

Por ejemplo, el módulo de clonación de datos de CA Test Data Manager extrae conjuntos de datos de pruebas pequeños y coherentes de varios sistemas de desarrollo y producción interrelacionados, lo que sustituye al lento y caro proceso de copiar y mover bases de datos enormes y complejas. La capacidad de extraer, copiar y proporcionar únicamente los datos necesarios se traduce en que las organizaciones ya no necesitan mantener numerosas copias de tamaño real de las bases de datos de producción.

Al disponer de un almacén de datos centralizado y poder clonarlos, resultará aún más sencillo eliminar las dependencias de datos entre equipos, con lo que se separará el aprovisionamiento y el consumo de los datos. Esto significa que los datos se pueden clonar y entregar a varios equipos en paralelo; así, se eliminan los retrasos que se pasan esperando a que los datos de fases previas estén disponibles y se evita que los equipos se perjudiquen los unos a los otros cuando introduzcan algún cambio en los datos.

Al modelar y almacenar centralmente los datos como objetos modificables y reutilizables, también es posible reproducir con facilidad errores y escenarios interesantes. Una vez que se hayan detallado los datos explícitamente en los informes de errores, la tecnología de clonación rápida y flexible posibilita la repetición de pruebas infrecuentes y complejas, sin que los datos se agoten. Se trata de una función especialmente valiosa a la hora de realizar una actualización de datos, ya que significa que los datos no se tienen que combinar y que los conjuntos de datos interesantes no se perderán.

Limitaciones de hardware

En combinación con un kit de herramientas de virtualización, la generación de datos sintéticos también puede contribuir a resolver las limitaciones del sistema y de hardware. Es posible simular las capas de mensajes utilizando los metadatos para generar perfiles precisos y mensajes de servicio realistas (incluidos archivos SOAP, REST y MQ, así como archivos sin formato). En este caso, existe un motor de generación automatizada de datos que subyace a la máquina virtual, a fin de crear respuestas a mensajes realistas, como pares solicitud-respuesta.

La virtualización de máquinas completas también trae consigo la posibilidad de crear varios entornos de desarrollo. Es decir, los equipos podrán trabajar en los entornos, incluso cuando haya componentes interdependientes que no estén disponibles, con lo que se evitarán los retrasos derivados de fases previas; todo ello a la vez que también se pueden virtualizar hardware y sistemas heredados de gran coste para efectuar las pruebas.

Proceso

Reutilización y desarrollo en paralelo

Además de los problemas relativos al entorno, la capacidad de detectar y clonar datos según atributos específicos de estos también da respuesta a otra preocupación central de la TDM, que es cómo se van a aprovisionar los datos de forma eficiente a equipos de pruebas o evaluadores individuales. Permite solicitar, compartir y reutilizar los datos en paralelo y a petición.

Al poder acceder a un portal de datos de pruebas a petición centralizado y basado en web, como el que se proporciona con CA Test Data Manager, por ejemplo, los evaluadores y desarrolladores pueden solicitar exactamente los datos que precisan para la labor que deben desempeñar. Cuando se envían criterios específicos (es decir, atributos de los datos de pruebas), el portal envía una tarea al motor de lotes, que encontrará los datos adecuados en los sistemas de back-end o clonará aquellos y los transferirá. Así se elimina la necesidad de buscar los datos o crearlos manualmente, con lo que se reduce de forma considerable el tiempo que se tarda en procesar las solicitudes de datos.

Cuanto más estandarizadas sean las preguntas del formulario, mejor, ya que así a los equipos les será más sencillo reutilizar el trabajo de los demás. Por ejemplo, si cuenta con datos sintéticos creados con anterioridad, podría parametrizar los campos de entrada y publicarlos mediante listas desplegables en el portal. De este modo, todos sus empleados podrían solicitar los datos, aunque el caso de prueba sea distinto. Además de los datos de pruebas, los marcos de trabajo de creación de datos, las pruebas unitarias, los activos virtuales y los scripts de automatización se pueden almacenar y utilizar como cimientos para iniciativas futuras.

Control de versiones

Un potente control de versiones ofrece el paralelismo necesario para desarrollar continuamente nuevos sistemas, a tiempo y sin superar el presupuesto. Por ejemplo, el almacén de datos de pruebas de CA Test Data Manager permite que un equipo copie datos de un repositorio, lo que en realidad hace que se "herede", junto con puntos desde los que revertirlos a versiones anteriores. De este modo, se contempla la evolución de los datos a lo largo de diversas versiones, ya que bloquea los datos para un equipo concreto y, al mismo tiempo, permite que se reviertan o actualicen con facilidad, y se reconcilien con las distintas versiones. Cuando se efectúa un cambio en alguna parte, se extiende a versiones previas y posteriores, mientras que la original permanece intacta. Supongamos que un equipo necesita agregar una columna en una base de datos completa: con CA Test Data Manager, si cuentan con objetos principales no modificables, podrán encontrar todos los elementos secundarios enlazados y valores predeterminados definidos, o bien generar datos mediante secuencias o funciones estándares.

Personas

Por último, con frecuencia, los cambios en las estructuras de los equipos pueden respaldar estas mejoras de tecnología y procesos, así como contribuir a dar respuesta a los puntos débiles relativos a los entornos que se han expuesto en el documento. La centralización de la TDM bajo el amparo de un equipo dedicado se traduce en contar con un recurso de almacenamiento, gestión y aprovisionamiento de los datos centralizado que pueda satisfacer las necesidades de la empresa con mayor eficiencia. Este equipo podría ocuparse del aprovisionamiento de los datos, así como de su gestión, y podría ser responsable de crear nuevos datos y perfiles de estos últimos en caso necesario.

Esto no solo contribuye a evitar las restricciones de dependencias de datos entre los equipos, sino que también significa que las solicitudes de datos y los informes de errores se pueden unificar en un solo ámbito. Por lo tanto, es posible aplicar controles de calidad, mientras que la propiedad de los datos se puede centralizar en el equipo de seguridad de TI. La creación de formularios dinámica que ofrece el portal de datos de pruebas a petición de CA Test Data Manager respalda esta ventaja, ya que no se limita a los derechos de acceso basados en roles, sino que solo aprovisiona los datos confidenciales al personal autorizado que los solicite.

Sección 4

Referencias

1 Las desventajas de utilizar datos de producción en entornos que no sean de producción se han expuesto en otros documentos. Consulte, por ejemplo: Huw Price, Reduce Time to Market with Test Data Management y How better Test Data Management is the only way to drive Continuous Delivery.

2 Jacek Becla y Daniel L. Wang, Lessons Learned from managing a Petabyte, P. 4. Consultado el 19/02/2015 en http://www.slac.stanford.edu/BFROOT/www/Public/Computing/Databases/proceedings/

3 Lessons Learned from managing a Petabyte

4 http://www.computerweekly.com/feature/Meeting-the-demand-for-data-storage

5 http://www.ponemon.org/blog/ponemon-institute-releases-2014-cost-of-data-breach-global-analysis

6 http://benderrbt.com/Bender-Requirements%20Based%20Testing%20Process%20Overview.pdf

7 http://www.softwaretestingclass.com/why-testing-should-start-early-in-software-development-lifecycle/

Sección 5

La ventaja de CA Technologies

CA Technologies (NASDAQ: CA) proporciona soluciones de gestión de TI que ayudan a los clientes a gestionar y proteger entornos de TI complejos para dar soporte a servicios de negocio ágiles. Las organizaciones utilizan el software y las soluciones de SaaS de CA Technologies para acelerar la innovación, transformar infraestructuras y proteger datos e identidades desde el centro de datos hasta la nube. En CA Technologies estamos comprometidos a asegurarnos de que nuestros clientes logren los resultados y el valor de negocio que esperan mediante el uso de nuestra tecnología. Para obtener más información sobre nuestros programas de éxito de clientes, visite **ca.com/customer-success**. Para obtener más información sobre CA Technologies, vaya a **ca.com/es**.

Sección 6

Acerca del autor

Con una carrera profesional que alcanza casi los 30 años, Huw Price ha sido el arquitecto técnico principal de numerosas empresas de software de EE. UU. y Europa, y ha proporcionado asistencia en el diseño arquitectónico de alto nivel a bancos multinacionales y a los principales proveedores de servicios y de servicios sanitarios. Recibió el premio "IT Director of the Year 2010" (Director de TI de 2010) de mano de QA Guild. Huw se ha especializado a lo largo de los años en las herramientas de automatización de pruebas y ha lanzado numerosos productos innovadores que han reestructurado el modelo de pruebas que se usa en el sector del software. En la actualidad, participa como ponente en eventos reconocidos internacionalmente y su trabajo se ha publicado en varias revistas como Professional Tester, CIO Magazine y otras publicaciones técnicas.

CA Technologies adquirió la última empresa de Huw, Grid-Tools, en junio de 2015. Durante casi una década, ya se había dedicado a redefinir la manera en que las grandes organizaciones enfocan su estrategia para realizar pruebas. Con el liderazgo y el enfoque visionario de Huw, la compañía ha introducido una sólida metodología de pruebas centrada en los datos, lo que ha derivado en nuevos conceptos acuñados por Huw, como "objetos de datos", "transmisión de datos" y "almacén central de datos de pruebas".

Comuníquese con CA Technologies en ca.com/es

CA Technologies (NASDAQ: CA) crea software que impulsa la transformación de las empresas y les permite aprovechar las oportunidades que brinda la economía de las aplicaciones. El software se encuentra en el corazón de cada empresa, sea cual sea su sector. Desde la planificación hasta la gestión y la seguridad, pasando por el desarrollo, CA trabaja con empresas de todo el mundo para cambiar la forma en que vivimos, realizamos transacciones y nos comunicamos, ya sea a través de la nube pública, la nube privada, plataformas móviles, entornos de mainframe o entornos distribuidos. Para obtener más información, visite **ca.com/es**.

Copyright © 2016 CA, Inc. Todos los derechos reservados. Todas las marcas a las que se hace referencia en este documento pertenecen a sus respectivas empresas. Este documento no otorga garantía alguna y se ofrece únicamente con fines informativos. Las descripciones de funciones pueden ser exclusivas de los clientes aquí presentados y el rendimiento real de los productos puede variar.

CS200-205127