Ulf T. Mattsson es Director general de tecnología (CTO) de Protegrity. Es el creador de la arquitectura inicial de la tecnología de seguridad para la base de datos de Protegrity, en virtud de la cual la empresa posee diversas patentes. Su amplia experiencia en el sector de la TI y la seguridad incluye 20 años en IBM, donde se desempeñó como gerente de desarrollo de software y consultor para la organización de investigación y desarrollo en las áreas de arquitectura y seguridad de TI.

Acortando la brecha entre el acceso y la seguridad en grandes conjuntos de datos

Las organizaciones fallan a la hora de proteger de verdad su información confidencial en entornos de grandes conjuntos de datos. Los analistas de datos necesitan acceder a los datos para realizar un análisis significativo de manera eficaz y obtener un retorno sobre la inversión (ROI); y la seguridad de datos tradicional en general ha limitado ese acceso. Como resultado, se ha generado una enorme cantidad de violaciones a los datos y ha disminuido la privacidad, seguido de elevadas multas y de la desintegración de la confianza del público. Es esencial garantizar la privacidad de las personas y una seguridad adecuada sin perder la capacidad de usar los datos y permitir que las organizaciones utilicen la información confidencial de manera responsable en su proprio beneficio.

ACCESO A (GRANDES) CONJUNTOS DE DATOS

La plataforma para grandes conjuntos de datos Hadoop se usa en este artículo para ejemplificar problemas de seguridad frecuentes, así como sus soluciones. Hadoop es la plataforma dominante para grandes conjuntos de datos, utilizada por una comunidad global, que carece de la seguridad de datos necesaria. Hadoop proporciona una plataforma de procesamiento masivamente paralela[1] diseñada para acceder a enormes cantidades de datos y experimentación, a fin de obtener nueva información a través del análisis y la comparación de más información de lo que solía ser posible o práctico previamente.

> "El cumplimiento no es garantía de seguridad."

Los datos fluyen más rápido, con mayor variedad, volumen y niveles de veracidad, y se pueden procesar de manera eficiente al acceder en forma simultánea a divisiones de datos entre cientos o miles de nodos de datos en un clúster. Además, los datos se guardan por períodos mucho más prolongados que en las bases de datos o sistemas de gestión de bases de datos relacionales (RDBMS), dado que el almacenamiento es más económico y el contexto histórico forma parte del diseño.

UNA FALSA SENSACIÓN DE SEGURIDAD

Si el objetivo principal de la plataforma Hadoop es el acceso a los datos, su seguridad se considera tradicionalmente como su antítesis. Siempre ha existido un tironeo entre ambas visiones sobre la base del riesgo, el equilibrio del desempeño operacional y la privacidad, pero el tema se magnifica en forma exponencial en Hadoop (**figura 1**).

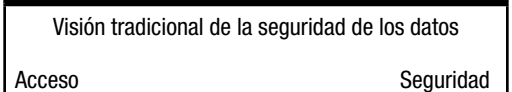
Por ejemplo, se pueden usar millones de registros personales para el análisis y para darle sentido a los datos, pero la privacidad de todas las personas se puede ver afectada en forma severa con una sola instancia de violación de datos. El riesgo involucrado es demasiado elevado para aceptar un nivel de seguridad bajo, pero obstruir el desempeño u obstaculizar la obtención de datos puede socavar la plataforma.

A pesar de la percepción de que el carácter confidencial de los datos representa un obstáculo para su acceso, dichos datos confidenciales en plataformas de grandes conjuntos de datos todavía requieren seguridad para cumplir con diversas reglamentaciones y leyes[2], de manera similar a cualquier otra plataforma de datos. Por lo tanto, la seguridad de datos en Hadoop se suele abordar desde la perspectiva del cumplimiento regulatorio.

Se podría suponer que esto contribuye a garantizar una seguridad máxima de los datos y un nivel de riesgo mínimo y en efecto, en cierta medida a obligar a las organizaciones a proteger sus datos. Sin embargo, como la seguridad se considera obstructiva al acceso de datos y por ende, al desempeño operacional, las reglamentaciones en realidad funcionan como guía para la implementación del menor nivel posible de seguridad necesario para cumplirlas. El cumplimiento no es garantía de seguridad.

Desde luego, las organizaciones desean proteger sus datos y la privacidad de sus clientes, pero el acceso, la información y el desempeño son esenciales. Para alcanzar el máximo nivel de acceso y seguridad de los datos, se debe acortar la brecha entre ambos. Entonces, ¿cuál es la mejor manera de lograr este equilibrio?

Figura 1 - Visión tradicional de la seguridad de los datos



Fuente: Ulf T. Mattsson. Reimpreso con autorización.

HERRAMIENTAS DE SEGURIDAD DE DATOS

Al momento de la elaboración de este artículo, Hadoop no cuenta con seguridad de datos nativa, aunque muchos proveedores tanto de Hadoop como de otros elementos para la seguridad de datos ofrecen soluciones complementarias.[3] Estas soluciones por lo general se basan en el control de acceso y/o la autenticación, ya que proporcionan un nivel inicial de seguridad con niveles de acceso relativamente elevados.

Control de acceso y autenticación

La implementación más común de autenticación en Hadoop es Kerberos.[4] En el control de acceso y la autenticación, los datos confidenciales están legibles durante las tareas laborales, al igual que en tránsito y almacenados. Además, ni el control de acceso ni la autenticación aportan demasiada protección respecto a los usuarios privilegiados, como los desarrolladores o administradores del sistema, que pueden eludirlos fácilmente para usar dichos datos en forma inapropiada. Por esta razón, muchas reglamentaciones, como la Norma de Seguridad de Datos de la Industria de las Tarjetas de Pago (PCI DSS)[5] y la Ley de Portabilidad y Responsabilidad de Seguros Médicos de EE. UU. (HIPAA),[6] exigen un nivel de seguridad más elevado para su cumplimiento.

Encriptación global

A partir de una base de controles de acceso y/o autenticación, incorporar encriptación global para todo un volumen o para todo un disco suele ser la primera opción para la seguridad de datos real en Hadoop. Este método es el de menor dificultad de implementación, que además cumple con las reglamentaciones. Los datos se encuentran protegidos cuando están almacenados (para archivo o disposición) y la encriptación por lo general es transparente a los usuarios y procesos autorizados. El resultado permite contar con niveles de acceso elevados, pero los datos en tránsito, en uso o bajo análisis están siempre legibles y los usuarios privilegiados de todos modos pueden acceder a los datos confidenciales. Este método solo ofrece protección ante el robo físico.

Encriptación fina

Si se agrega encriptación robusta para las columnas o campos, se logra un mayor nivel de seguridad, se protegen los datos almacenados, en tránsito y respecto a usuarios privilegiados, pero es necesario que se revelen los datos en forma visible (desencriptados) para desempeñar las funciones propias del trabajo, incluido el análisis, dado que los datos encriptados son ilegibles para usuarios y procesos.

La encriptación con preservación del formato conserva la capacidad para usuarios y aplicaciones de leer los datos protegidos, pero es uno de los procesos de encriptación más lentos.

La implementación de cualquiera de estos métodos puede afectar de manera significativa el desempeño, incluso con los procesos de encriptación/desencriptación más rápidos disponibles, lo que anula muchas de las ventajas de la plataforma Hadoop. Dado que el acceso es primordial, estos métodos inclinan la balanza excesivamente en favor de la seguridad como para resultar viables.

Algunos proveedores ofrecen un sistema de archivos virtual superior al Sistema de archivos distribuido Hadoop (HDFS), con encriptación dinámica de datos basada en roles. Aunque esta opción ofrece cierto nivel de seguridad de los datos en uso, no hace nada para proteger los datos bajo análisis o respecto de los usuarios privilegiados, que pueden acceder al sistema operativo (SO) y a niveles inferiores al virtual y llegar a los datos no encriptados.

Enmascaramiento de datos

El enmascaramiento preserva el tipo y la extensión de los datos estructurados, y los reemplaza por un valor inerte y sin sentido. Debido a que los datos enmascarados se ven y actúan como los originales, son legibles para los usuarios y los procesos.

El enmascaramiento de datos estáticos (SDM) reemplaza permanentemente los valores de información confidencial con datos inertes. SDM se suele usar para desempeñar funciones de trabajo preservando una cantidad suficiente de los datos originales o haciendo que los datos no puedan identificarse. Protege los datos almacenados, en uso, en tránsito, bajo análisis y respecto a usuarios privilegiados. No obstante, en caso que se vuelvan a necesitar alguna vez los datos en texto legible (es decir, para realizar operaciones de marketing o en escenarios de salud), son irrecuperables. Por lo tanto, se utiliza SDM en entornos de prueba/desarrollo en los que se necesitan datos que se ven y actúan como datos reales para las pruebas, pero los datos confidenciales no se exponen a los desarrolladores o administradores del sistema. No se suele usar para el acceso de datos en un entorno Hadoop de producción. Según los algoritmos de enmascaramiento utilizados y qué datos se reemplazan, es posible que los datos de SDM estén sujetos a la inferencia y puedan volver a identificarse si se combinan con otras fuentes de datos.

El enmascaramiento dinámico de datos (DDM) realiza el enmascaramiento "sobre la marcha". Cuando se solicitan datos confidenciales, se establecen referencias de la política y se recuperan los datos enmascarados que el usuario o proceso no está autorizado a ver sin encriptación, sobre la

base de la función del usuario o proceso. De manera similar a la encriptación de datos dinámicos y al control de acceso, DDM no proporciona seguridad a los datos alamcenados o en tránsito, y escasa protección de los usuarios privilegiados. Los valores enmascarados dinámicamente también pueden ser problemáticos para trabajar en escenarios analíticos de producción, según el algoritmo/método utilizado[7].

Tokenización

La tokenización también reemplaza el texto legible por un valor inerte aleatorio del mismo tipo y extensión del dato, pero el proceso se puede revertir. Esto se logra través del uso de tablas de tokens, en lugar de un algoritmo de encriptación. En la tokenización sin bóveda, se reemplazan pequeños bloques de los datos originales con valores emparejados aleatorios de las tablas de tokens que se superponen entre bloques. Una vez que se ha tokenizado todo el valor, se vuelve a ejecutar el proceso para eliminar cualquier patrón de la transformación.

Sin embargo, debido a que el valor resultante depende del valor ingresado, puede mantenerse de todos modos una relación uno a uno con los datos originales y por lo tanto, se pueden utilizar los datos tokenizados en el análisis como reemplazo del texto legible. Además, se pueden preservar partes de los datos en texto legible o se los puede "filtrar" usando el token, lo que resulta de especial utilidad en casos en que solo se requiere parte de los datos originales para realizar una tarea.

La tokenización también permite cierta flexibilidad en el nivel de privilegios de seguridad de los datos, ya que se puede otorgar la autorización campo por campo o por campo parcial. Los datos se protegen en todos los estados: almacenados, en uso, en tránsito y en análisis.

ACORTAR LA BRECHA

Al comparar los métodos de la seguridad de datos finos (figura 2), resulta evidente que la tokenización ofrece los mayores niveles de accesibilidad y seguridad. Los valores de tokens aleatorizados no tienen valor alguno para un potencial delincuente, ya que solo aquellos autorizados para acceder a la tabla y al proceso de tokens pueden esperar recuperar el valor original de la tabla. La capacidad de utilizar los valores tokenizados en el análisis representa una mayor seguridad y eficiencia, ya que los datos permanecen protegidos y no requieren procesamiento adicional para desprotegerlos o destokenizarlos.

La capacidad de extraer de manera segura el valor de los datos confidenciales que no permitan identificación es la clave para acortar la brecha entre la privacidad y el acceso. Los datos protegidos permanecen utilizables para la mayoría de los usuarios y procesos, y solo aquellos con privilegios otorgados de conformidad con la política de seguridad de los datos pueden acceder a los datos confidenciales en forma legible.

METODOLOGÍA DE SEGURIDAD DE LOS DATOS

La tecnología para la seguridad de datos no es en sí misma suficiente para asegurar un equilibrio optimizado entre acceso y seguridad. Después de todo, cualquier sistema es solo tan fuerte como su eslabón más débil. En el área de la seguridad de datos, ese eslabón suele ser el componente humano. Como tal, se puede utilizar una metodología clara y concisa para contribuir a optimizar los procesos de seguridad de datos y minimizar el impacto en las operaciones comerciales (**figura 3**).

Figura 2 — Comparación de métodos finos de seguridad de datos

Métodos de seguridad de datos	Desempeño	Almacenamiento	Seguridad	Transparencia
Sistema sin protección de datos	●	●	○	●
Monitoreo + bloqueo + ofuscación	◔	●	◔	◔
Encriptación de preservación de tipo de datos	◒	●	◒	◔
Encriptación robusta	◔	◔	●	◔
Tokenización sin bóveda	●	●	●	◔
Hashing	●	◔	●	○
Anonimización	●	●	●	○

Peor ○ ◔ ◒ ◔ ● Mejor

Fuente: Ulf T. Mattsson. Reimpreso con autorización.

Figura 3 — Metodología de seguridad de los datos	
Clasificación	Determinar qué datos son confidenciales para la organización, ya sea para cumplimiento regulatorio y/o internamente.
Descubrimiento	Determinar la ubicación de los datos confidenciales, la manera en que fluyen, quiénes pueden acceder a ellos, el desempeño y otros requisitos de seguridad.
Seguridad	Aplicar los métodos de seguridad de los datos que permiten alcanzar los requisitos de descubrimiento de manera más eficaz, y proteger los datos según la criticidad determinada por su clasificación.
Cumplimiento	Diseñar e implementar una política de seguridad de los datos destinada a revelar los datos confidenciales solo a los usuarios autorizados, según la menor cantidad posible de información necesaria para desempeñar las tareas laborales (principio del menor privilegio).
Monitoreo	Asegurar el monitoreo continuo, altamente granular, de cualquier intento de acceder a datos confidenciales. El monitoreo es la única defensa contra el abuso de datos por parte de los usuarios autorizados.
Fuente: Ulf T. Mattsson. Reimpreso con autorización	

Clasificación
La primera consideración de la implementación de seguridad de datos debe ser una clasificación clara de qué datos se consideran confidenciales, según las reglamentaciones externas y/o las directivas internas de seguridad. Esto puede incluir cualquier dato, desde información personal a resultados de análisis de operaciones internas.

Descubrimiento
Determinar dónde se encuentran ubicados los datos confidenciales, al igual que sus fuentes y dónde se utilizan, es el próximo paso de una metodología de seguridad de datos básica. Es posible que un tipo de dato específico requiera diferentes niveles de protección en distintas partes del sistema. Comprender el flujo de datos es esencial para protegerlos.

Además, Hadoop no debe considerarse un elemento aislado externo a la empresa. El procesamiento analítico en Hadoop por lo general solo forma parte del proceso global, desde fuentes de datos a Hadoop, a las bases de datos, hasta las plataformas de análisis más detallado. La implementación de la seguridad de datos para toda la empresa puede proteger los datos en distintas plataformas de manera más consistente, lo que minimiza las brechas y puntos de fuga.

Seguridad
Luego, la selección de los métodos de seguridad que se adecuen mejor al riesgo, tipo de dato y caso de uso de cada clasificación de datos confidenciales, o elementos de dato, asegura la implementación de la solución más eficaz para todos los datos confidenciales. Por ejemplo, aunque la tokenización sin bóveda ofrece un acceso y seguridad sin precedentes para los datos estructurados, como los números y nombres de tarjetas de crédito, se puede utilizar la encriptación para datos no analíticos no estructurados, como las imágenes y otros archivos de los medios.

También es importante proteger los datos lo antes posible, tanto en la implementación de Hadoop como en la adquisición/creación de datos. Esto contribuye a limitar la posible exposición de los datos confidenciales no encriptados.

Cumplimiento
Diseñar una política de seguridad de los datos basada en el principio del menor privilegio (es decir, revelar la menor cantidad posible de datos confidenciales no encriptados que resulte necesaria para desempeñar las tareas laborales). Esto se puede lograr a través de la creación de roles de políticas que determinen quién tiene acceso o quién no, según qué grupo es menor. Un abordaje moderno del control de acceso puede permitir que el usuario vea diferentes perspectivas de un campo de datos específico y así, exponga mayor o menor cantidad del contenido confidencial de ese campo de datos.

Es muy importante asignar la responsabilidad de la administración de la política de seguridad de los datos y su cumplimiento al equipo de seguridad. Los límites desdibujados entre la seguridad y la gestión de datos que existen en muchas organizaciones conducen a usos indebidos de datos confidenciales por parte de los usuarios privilegiados, que pueden ser de gravedad. Esta separación de tareas previene la mayoría de los usos indebidos, al crear un control automatizado sólido y asignar responsabilidad sobre el acceso a datos no encriptados.

Monitoreo
Al igual que con cualquier solución de seguridad de datos, se debe utilizar el monitoreo extensivo de datos confidenciales en Hadoop. Incluso aunque se implemente una seguridad de datos adecuada, el monitoreo inteligente puede agregar un nivel de control de acceso de datos basado en el contexto, a fin de asegurar de que los usuarios autorizados no utilicen los datos en forma indebida.

¿Qué distingue un usuario autorizado de uno privilegiado? Los usuarios privilegiados por lo general son miembros del equipo de TI con acceso privilegiado a la plataforma de datos. Estos usuarios pueden incluir a los administradores del sistema o a analistas que tienen acceso casi sin limitaciones a los sistemas a los fines de realizar tareas de mantenimiento y desarrollo. Los usuarios autorizados son aquellos a los que el equipo de seguridad les ha concedido acceso para visualizar datos confidenciales.

El monitoreo altamente granular de datos confidenciales es esencial para asegurar la detección temprana de las amenazas externas e internas.

CONCLUSIÓN

Cumplir estas buenas prácticas permitiría que las organizaciones extraigan el valor de los datos confidenciales de manera segura y que adopten plataformas de grandes conjuntos de datos con mucho menor riesgo de violaciones a su seguridad. Además, la protección y la preservación de la privacidad de los clientes y demás personas contribuye a proteger el nombre y la reputación de la organización.

Es posible obtener información del análisis profundo de datos, junto con una verdadera seguridad para los mismos. Con el tiempo y los conocimientos necesarios, cada vez más organizaciones lo lograrán.

NOTAS FINALES

[1] The Apache Software Foundation, *http://hadoop.apache.org*. La biblioteca Apache de software de Hadoop Apache es un marco que permite el procesamiento distribuido de grandes conjuntos de datos en clústers de computadoras que utilizan modelos simples de programación. Está diseñada para permitir el escalamiento desde servidores individuales a miles de máquinas, cada una con computación y almacenamiento locales. En lugar de depender del hardware para ofrecer alta disponibilidad, la biblioteca en sí está diseñada para detectar y manejar fallas a nivel de la aplicación y así, prestar un servicio de alta disponibilidad sobre un clúster de computadoras que en forma individual puede presentar una tendencia a las fallas.

[2] Las reglamentaciones de aplicación frecuente incluyen la Ley de Portabilidad y Responsabilidad de Seguros Médicos de EE. UU. (HIPAA), la Norma de Seguridad de Datos de la Industria de las Tarjetas de Pago (PCI DSS), La Ley Sarbanes-Oxley de EE. UU. y las leyes estatales o nacionales relativas a la privacidad de los datos.

[3] Estos proveedores de soluciones incluyen a Cloudera, Gazzang, IBM, Intel (código abierto), MIT (código abierto), Protegrity y Zettaset, que ofrecen una o más de las siguientes soluciones: control de acceso, autenticación, encriptación de volumen, encriptación de campo/columna, enmascaramiento, tokenización y/o monitoreo.

[4] Massachusetts Institute of Technology (MIT), Estados Unidos, *http://web.mit.edu/kerberos/*. Kerberos, desarrollado originalmente por el Proyecto Athena del MIT, es un protocolo de autenticación de red generalizado. Está diseñado para ofrecer autenticación sólida para aplicaciones de servidores de clientes mediante el uso de criptografía de clave secreta.

[5] PCI Security Standards Council, *www.pcisecuritystandards.org*. Las normas PCI DSS aportan lineamientos y regulan la protección de los datos de las tarjetas de pago, incluido el número de cuenta primario (PAN), los nombres, el número de identificación personal (PIN) y otros componentes relacionados con su procesamiento.

[6] Departamento de Salud y Servicios Humanos de los Estados Unidos, *www.hhs.gov/ocr/privacy*. La Reglamentación de seguridad HIPAA especifica una serie de medidas administrativas, físicas y técnicas para las entidades cubiertas y sus socios comerciales, destinadas a asegurar la confidencialidad, integridad y disponibilidad de la información electrónica de salud protegida.

[7] Los valores enmascarados dinámicamente suelen mezclarse en forma independiente, lo que puede disminuir drásticamente la utilidad de los datos en el análisis relacional, debido a que los campos de referencia ya no están alineados. Además, los valores pueden tener coincidencia cruzada o falsa coincidencia, si están truncados o se reemplazan parcialmente con datos no aleatorios (como los hashes). El tema radica en el hecho de que los valores enmascarados por lo general no son generados sino referenciados dinámicamente, como un subconjunto enmascarado separado de los datos originales.