

OFUSCACIÓN DEL CONTENIDO DE USUARIO EN ARCHIVOS DE DATOS ESTRUCTURADOS DE USUARIO

Antecedentes de la invención

Diversas aplicaciones de productividad de usuario permiten la introducción de datos y el análisis del contenido de usuario. Estas aplicaciones pueden proporcionar la creación, edición y análisis de contenido utilizando hojas de cálculo, presentaciones, documentos de texto, documentos multimedia combinados, formatos de mensajería u otros formatos de contenido de usuario. Entre este contenido de usuario, la información textual, alfanumérica u otra basada en caracteres puede incluir datos confidenciales que los usuarios u organizaciones pueden no desear incluir en trabajos publicados o distribuidos. Por ejemplo, una hoja de cálculo puede incluir números de seguro social (SSN, por sus siglas en inglés), información de tarjeta de crédito, identificadores de atención médica u otra información. Aunque, el usuario que ingresa estos datos o el contenido de usuario puede tener autorización para visualizar los datos confidenciales, otras entidades o terminales de distribución pueden no tener tal autorización.

Las técnicas de gestión y protección de la información pueden denominarse protección contra pérdida de datos (DLP, por sus siglas en inglés) que intenta evitar la apropiación indebida y la designación incorrecta de estos datos confidenciales. En ciertos formatos de contenido o tipos de contenido, como aquellos incluidos en hojas de cálculo, presentaciones con base en diapositivas y aplicaciones de diagramación gráfica, el contenido de usuario puede incluirse en varias celdas, objetos u otras entidades de datos estructurados o semiestructurados. Además, los datos confidenciales pueden dividirse entre más de una entidad de datos. Pueden surgir dificultades al intentar identificar y protegerse contra la pérdida de datos confidenciales cuando tales documentos incluyen datos confidenciales.

Breve descripción de la invención

En el presente documento se proporcionan sistemas, métodos y *software* para estructuras de ofuscación de datos para aplicaciones de usuario. Un método ejemplar incluye proporcionar contenido de usuario a un servicio de clasificación configurado para procesar el contenido de usuario para clasificar porciones del contenido de usuario como contenido confidencial, y recibir del servicio de clasificación las indicaciones del contenido de usuario que contiene el contenido confidencial. El método incluye presentar indicaciones gráficas en una interfaz de usuario (UI, por sus siglas en inglés) a la aplicación de usuario que señalan que el contenido de usuario incluye el contenido confidencial, y presentan opciones de ofuscación en la interfaz de usuario para ocultar el contenido confidencial dentro de al menos una porción seleccionada de entre el contenido de usuario. En respuesta a una selección del usuario de al menos una de las opciones de ofuscación, el método incluye reemplazar el contenido de usuario asociado con contenido ofuscado que mantiene un esquema de datos del contenido de usuario asociado.

Se proporciona esta “breve descripción de la invención” para introducir una selección de conceptos de una manera simplificada que además se describen más adelante en la “descripción detallada de la invención”. Esta “breve descripción de la invención” no busca identificar factores clave o características

esenciales de la materia objeto reclamada, ni busca usarse para limitar el alcance de la materia objeto reclamada.

Breve descripción de los dibujos

Muchos aspectos de la divulgación pueden entenderse mejor con referencia a los siguientes dibujos. Aunque se describen varias implementaciones con relación a estos dibujos, la divulgación no se limita a las implementaciones divulgadas en el presente documento. Por el contrario, la intención es cubrir todas las alternativas, modificaciones y equivalentes.

La figura 1 ilustra un entorno de protección contra pérdida de datos en un ejemplo.

La figura 2 ilustra elementos de un entorno de protección contra pérdida de datos en un ejemplo.

La figura 3 ilustra elementos de un entorno de protección contra pérdida de datos en un ejemplo.

La figura 4 ilustra las operaciones de los entornos de protección contra pérdida de datos en un ejemplo.

La figura 5 ilustra las operaciones de los entornos de protección contra pérdida de datos en un ejemplo.

La figura 6 ilustra las operaciones de los entornos de protección contra pérdida de datos en un ejemplo.

La figura 7 ilustra las operaciones de los entornos de protección contra pérdida de datos en un ejemplo.

La figura 8 ilustra las operaciones de umbral de datos de los entornos de protección contra pérdida de datos en un ejemplo.

La figura 9 ilustra un sistema informático adecuado para implementar cualquiera de las arquitecturas, procesos, plataformas, servicios y casos operativos divulgados en el presente documento.

Descripción detallada de la invención

Las aplicaciones de productividad de usuario proporcionan la creación, edición y análisis de datos y contenido de usuario utilizando hojas de cálculo, diapositivas, elementos gráficos vectoriales, documentos, correos electrónicos, contenido de mensajería, bases de datos u otros formatos y tipos de datos de aplicaciones. Entre el contenido de usuario, puede incluirse información textual, alfanumérica u otra basada en caracteres. Por ejemplo, una hoja de cálculo puede incluir números de seguro social (SSN, por sus siglas en inglés), información de tarjeta de crédito, identificadores de atención médica, números de pasaporte u otra información. Aunque el usuario que ingresa estos datos o el contenido de usuario puede tener autorización para visualizar los datos confidenciales, otras entidades o terminales de distribución pueden no tener tal autorización. Pueden establecerse varias políticas de privacidad o reglas de privacidad de datos que indican qué tipos de datos o contenido de usuario son de naturaleza confidencial. Pueden incorporarse las medidas mejoradas de protección contra pérdida de datos (DLP, por sus siglas en inglés) abordadas en el presente documento para intentar evitar la apropiación indebida y la asignación incorrecta de estos datos confidenciales.

En ciertos formatos de contenido o tipos de contenido, como los incluidos en hojas de cálculo, presentaciones con base en diapositivas y aplicaciones de diagramación gráfica, el contenido de usuario puede incluirse en varias celdas, objetos u otras entidades de datos estructurados o semiestructurados. Además, los datos confidenciales pueden dividirse entre más de un elemento de datos o entrada. Los ejemplos en este documento proporcionan una identificación mejorada de datos confidenciales en archivos de datos de usuario que incluyen elementos de datos estructurados. Además, los ejemplos en este documento proporcionan interfaces de usuario mejoradas para alertar a los usuarios sobre datos confidenciales. Estos elementos de la interfaz de usuario pueden incluir la señalización de elementos de datos que contienen datos confidenciales individuales, así como umbrales para alertar durante la edición del contenido.

En una aplicación ejemplar que utiliza los elementos de datos estructurados, como una aplicación de hoja de cálculo, los datos pueden ingresarse en celdas que están dispuestas en columnas y filas. Cada celda puede contener datos de usuario o contenido de usuario y también puede incluir una o más expresiones que se utilizan para realizar cálculos, que pueden hacer referencia a datos ingresados por parte del usuario en una o más celdas. Otras aplicaciones de usuario, como las aplicaciones de presentación de diapositivas, pueden incluir contenido de usuario en más de una diapositiva, así como dentro de los objetos incluidos en estas diapositivas.

Ventajosamente, los ejemplos e implementaciones en este documento proporcionan operaciones y estructuras mejoradas para servicios de protección contra pérdida de datos. Estas operaciones y estructuras mejoradas tienen efectos técnicos de identificación más rápida de contenido confidencial dentro de documentos y especialmente para documentos estructurados, como hojas de cálculo, presentaciones, dibujos esquemáticos y similares. Además, varias aplicaciones pueden compartir un único servicio de clasificación que proporciona detección e identificación de contenido confidencial en archivos de datos de usuario en numerosas y variadas aplicaciones y plataformas de usuario final. Los procesos de anotación y ofuscación a nivel del usuario final también proporcionan ventajas significativas y efectos técnicos en las interfaces de usuario para las aplicaciones. Por ejemplo, a los usuarios se les pueden presentar anotaciones gráficas de contenido confidencial y cuadros de diálogo emergentes que presentan varias opciones de ofuscación u ocultación. También pueden establecerse varios umbrales de anotación mejorados para indicar dinámicamente el contenido confidencial a los usuarios para hacer más eficiente la edición de contenido y la ofuscación de datos confidenciales y que cumplan con varias políticas y reglas de protección contra pérdida de datos.

Como primer ejemplo de un entorno de protección contra pérdida de datos para una aplicación de usuario, se proporciona la figura 1. La figura 1 ilustra el entorno de protección contra pérdida de datos 100 en un ejemplo. El entorno 100 incluye la plataforma de usuario 110 y la plataforma de protección contra pérdida de datos 120. Los elementos de la figura 1 pueden comunicarse a través de uno o más enlaces de comunicación físicos o lógicos. En la figura 1, se muestran los enlaces 160-161. Sin embargo, debe entenderse que estos enlaces son solo ejemplares y pueden incluirse uno o más enlaces adicionales, que pueden incluir porciones inalámbricas, alámbricas, ópticas o lógicas.

Una estructura de protección contra pérdida de datos puede incluir una porción local para una aplicación de usuario específica y una porción compartida empleada en muchas aplicaciones. La plataforma de usuario 110 proporciona un entorno de aplicación para que un usuario interactúe con elementos de la aplicación de usuario 111 a través de la interfaz de usuario 112. Durante la interacción del usuario con la aplicación 111, puede realizarse la introducción de contenido y la manipulación de contenido. El módulo de protección contra pérdida de datos (DLP, por sus siglas en inglés) de aplicaciones 113 puede proporcionar porciones de la funcionalidad para la anotación y reemplazo de datos confidenciales dentro de la aplicación 111. El módulo de DLP de aplicaciones 113 es local para la plataforma de usuario 110 en este ejemplo, pero en su lugar podría estar separado o integrado en la aplicación 111. El módulo de DLP de aplicación 113 puede proporcionar la anotación y reemplazo de datos confidenciales para los usuarios y la aplicación 111. La plataforma de protección contra pérdida de datos 120 proporciona una porción compartida de una estructura de protección contra pérdida de datos y proporciona un servicio compartido de DLP 121 para que muchas aplicaciones lo compartan, como las aplicaciones 190 con la porción de DLP de ubicación asociada 193.

En funcionamiento, la aplicación 111 proporciona la interfaz de usuario 112 a través de la cual los usuarios pueden interactuar con la aplicación 111, como ingresar, editar y manipular de otro modo el contenido de usuario que puede cargarse a través de uno o más archivos de datos o ingresarse a través de la interfaz de usuario 112. En la figura 1, se muestra una hoja de cálculo con celdas dispuestas en filas y columnas. Como parte de la aplicación 111, se proporciona un servicio de protección contra pérdida de datos que identifica el contenido confidencial del usuario y permite a los usuarios reemplazar el contenido confidencial del usuario con texto o datos seguros. El contenido confidencial comprende contenido que podría tener cuestiones de privacidad, políticas/reglas de privacidad u otras propiedades para las cuales su difusión sería indeseada o no deseada. La pérdida de datos en este contexto se refiere a la difusión de datos privados o confidenciales a usuarios o terminales no autorizados.

Para identificar el contenido confidencial, la aplicación 111 proporciona el contenido de usuario en piezas o fragmentos del contenido de usuario a un servicio de protección contra pérdida de datos. En la figura 1, las porciones de contenido 140 se muestran con porciones de contenido individuales 141-145 que se proporcionan durante el transcurso del tiempo al servicio de DLP 121. Por lo general, la aplicación 111 puede procesar el contenido de usuario para dividir el contenido de usuario en las porciones durante los períodos de inactividad, como cuando uno o más subprocesos para el procesamiento relacionados con la aplicación 111 están inactivos o por debajo de los umbrales de actividad. Tal como se abordará en el presente documento, el contenido estructurado de usuario se transforma en una disposición “reducida” o no estructurada durante el proceso de división. Esta disposición no estructurada tiene varias ventajas para el procesamiento por el servicio de DLP 121.

El servicio de DLP 121 entonces procesa cada porción o “fragmento” del contenido de usuario individualmente para determinar si las porciones contienen contenido confidencial. Pueden introducirse varias reglas de clasificación 125, como esquemas de datos, patrones de datos o políticas/reglas de privacidad en el servicio 121 de DLP para la identificación de datos confidenciales. Después de que el

servicio DLP 121 analiza cada fragmento individual de contenido de usuario, los intervalos de ubicación de los datos confidenciales en el archivo de datos del usuario se determinan indicados para el servicio DLP de aplicación 113. Una función de correlación en el servicio de DLP de aplicación 113 determina una relación estructural entre los intervalos de fragmentos y la estructura del documento. Las indicaciones de los intervalos de ubicación, las longitudes de datos confidenciales y los tipos de datos confidenciales pueden proporcionarse a la aplicación 111, tal como se observa, por ejemplo, en las indicaciones de datos confidenciales 150. Los intervalos de ubicación indicados por el servicio de DLP 121 podrían no producir una ubicación exacta o específica entre los elementos estructurales del archivo de datos de usuario para el contenido confidencial. En estos casos, el servicio de DLP de aplicación 113 de la aplicación 111 puede emplear un proceso de correlación para determinar los elementos estructurales específicos que contienen los datos confidenciales.

Una vez que se determinan ubicaciones específicas, la aplicación 111 puede anotar los datos confidenciales dentro de la interfaz de usuario 112. Esta anotación puede incluir una marca o señalización global o individual de los datos confidenciales. Las anotaciones pueden comprender “consejos de política” presentados en una interfaz de usuario. A los usuarios se les puede presentar una o más opciones para ofuscar o de otro modo hacer que el contenido de usuario no sea identificable como el contenido confidencial original. Pueden establecerse varios umbrales para la notificación del contenido confidencial que se activan con respecto al conteo o cantidades de los datos confidenciales presentes en el archivo de datos de usuario.

En un ejemplo, el archivo de datos de usuario 114 incluye contenido 115, 116 y 117 en celdas particulares del archivo de datos de usuario 114, que podrían asociarse con una hoja de trabajo o página particular del libro de la hoja de cálculo. Pueden incluirse diversos contenidos en las celdas asociadas, y este contenido puede comprender datos potencialmente confidenciales, como los ejemplos observados en la figura 1 para SSN, números de teléfono y direcciones. Parte de este contenido puede cruzar límites estructurales en el archivo de datos de usuario, como abarcar varias celdas o abarcar múltiples objetos gráficos. Si el “fragmento” divide los datos en filas o agrupaciones de filas, entonces las representaciones reducidas (es decir, despojadas de cualquier contenido estructural) aún pueden identificar datos confidenciales dentro de una o más celdas.

Los elementos de cada plataforma de usuario 110 y plataforma de DLP 120 pueden incluir interfaces de comunicación, interfaces de red, sistemas de procesamiento, sistemas informáticos, microprocesadores, sistemas de almacenamiento, medios de almacenamiento o algunos otros dispositivos de procesamiento o sistemas de *software*, y pueden distribuirse entre múltiples dispositivos o en múltiples ubicaciones geográficas. Los ejemplos de elementos de cada una de las plataformas de usuario 110 y de DLP 120 pueden incluir *software* como un sistema operativo, aplicaciones, registros, interfaces, bases de datos, utilidades, controladores, *software* de red y otro *software* almacenado en un medio legible por computadora. Los elementos de cada una de las plataformas de usuario 110 y de DLP 120 pueden comprender una o más plataformas que estén alojadas por un sistema informático distribuido o servicio de computación en la nube. Los elementos de cada una de las plataformas de usuario 110 y de DLP 120

pueden comprender elementos de interfaz lógica, como interfaces definidas por *software* e interfaces de programación de aplicaciones (API, por sus siglas en inglés).

Los elementos de la plataforma de usuario 110 incluyen la aplicación 111, interfaz de usuario 112 y módulo de DLP de aplicación 113. En este ejemplo, la aplicación 111 comprende una aplicación de hoja de cálculo. Debe entenderse que la aplicación de usuario 111 puede comprender cualquier aplicación de usuario, como aplicaciones de productividad, aplicaciones de comunicación, aplicaciones de redes sociales, aplicaciones de juegos, aplicaciones móviles u otras aplicaciones. La interfaz de usuario 112 comprende elementos gráficos de interfaz de usuario que pueden producir una salida para mostrarse a un usuario y recibir información de parte de un usuario. La interfaz de usuario 112 puede comprender los elementos que se abordan a continuación en la figura 9 para el sistema de interfaz de usuario 908. El módulo de DLP de aplicación 113 comprende uno o más elementos de *software* configurados para dividir contenido para entregar a un servicio de clasificación, anotar datos indicados como confidenciales y ofuscar datos confidenciales, entre otras operaciones.

Los elementos de la plataforma de DLP 120 incluyen el servicio de DLP 121. El servicio de DLP 121 incluye una interfaz externa en forma de interfaz de programación de aplicaciones (API, por sus siglas en inglés) 122, aunque pueden emplearse otras interfaces. El servicio de DLP 121 también incluye el seguidor 123 y el servicio de clasificación 124, que se analizarán con más detalle posteriormente. La API 122 puede incluir una o más interfaces de usuario, como interfaces web, API, interfaces de terminal, interfaces de consola, interfaces de intérprete de órdenes de línea de comandos, interfaces de lenguaje de marcado extensible (XML, por sus siglas en inglés), entre otras. El seguidor 123 mantiene conteos o cantidades de datos confidenciales encontrados para un documento particular dentro de porciones reducidas de contenido estructurado de usuario, y también mantiene un registro de intervalos de ubicación dentro de las porciones reducidas de contenido estructurado de usuario que corresponden a ubicaciones de datos confidenciales dentro del contenido estructurado de usuario. El seguidor 123 también puede realizar un análisis de umbral para determinar cuándo se encuentran cantidades umbral de datos confidenciales y debe anotarse por el módulo de DLP de aplicación 113. Sin embargo, en otros ejemplos, las porciones de umbral/recuento del servicio de DLP 121 podrían incluirse en el módulo de DLP 113. El servicio de clasificación 124 analiza el contenido reducido de usuario para determinar la presencia de datos confidenciales, y puede emplear varias entradas que definen reglas y políticas para identificar los datos confidenciales. Los elementos del módulo de DLP de aplicación 113 y el servicio de DLP compartido 121 pueden configurarse en diferentes disposiciones o distribuciones mostradas en la figura 1, como cuando porciones del servicio de DLP compartido 121 se incluyen en el módulo de DLP de aplicación 113 o aplicación 111, entre otras configuraciones. En un ejemplo, las porciones del servicio de DLP compartido 121 comprenden una biblioteca de enlace dinámico (DLL, por sus siglas en inglés) incluida en la plataforma de usuario 110 para uso de la aplicación 111 y del módulo de DLP de aplicación 113.

Los enlaces 160-161, junto con otros enlaces no mostrados entre los elementos de la figura 1 para mayor claridad, pueden comprender uno o más enlaces de comunicación, como uno o más enlaces de red que comprenden enlaces de red inalámbricos o alámbricos. Los enlaces pueden comprender varias

interfaces de programación lógicas, físicas o de aplicaciones. Los enlaces de comunicación ejemplares pueden utilizar metal, vidrio, óptica, aire, espacio o algún otro material como medio de transporte. Los enlaces pueden utilizar varios protocolos de comunicación, como el protocolo de internet (IP, por sus siglas en inglés), Ethernet, híbrido de fibra coaxial (HFC, por sus siglas en inglés), redes ópticas síncronas (SONET, por sus siglas en inglés), modo de transferencia asíncrona (ATM, por sus siglas en inglés), multiplexión por división de tiempo (TDM, por sus siglas en inglés), conmutación de circuitos, señalización de comunicación, comunicaciones inalámbricas o algún otro formato de comunicación, que incluyen combinaciones, mejoras o variaciones de los mismos. Los enlaces pueden ser enlaces directos o pueden incluir redes, sistemas o dispositivos intermedios, y pueden incluir un enlace de red lógica transportado a través de múltiples enlaces físicos.

Para un análisis adicional de los elementos y operación del entorno 100, se presenta la figura 2. La figura 2 es un diagrama de bloques que ilustra la configuración ejemplar 200 del módulo de DLP de aplicación 113, que destaca las operaciones ejemplares del módulo de DLP de aplicación 113, entre otros elementos. En la figura 2, el módulo de DLP de aplicación 113 incluye el repartidor de contenido 211, el anotador 212, el mapeador 213 y el ofuscador 214. Cada uno de los elementos 211-214 puede comprender módulos de *software* empleados por el módulo de DLP de aplicación 113 para operar tal como se describe posteriormente.

En operación, el contenido de usuario se proporciona al módulo de DLP de aplicación 113, como un archivo de hoja de cálculo o libro tal como se observa en la figura 1 para el archivo de datos de usuario 114. Este archivo de datos de usuario puede organizarse en un formato estructurado o semiestructurado, como celdas organizadas por filas y columnas para un ejemplo de hoja de cálculo. En cambio, pueden emplearse otros formatos de datos, como presentaciones de diapositivas con páginas/diapositivas y numerosos objetos gráficos individuales, programas de dibujo vectorial con varios objetos en varias páginas, documentos de procesamiento de texto con varios objetos (tablas, cuadros de texto, imágenes), bases de datos, contenido de página web u otros formatos, que incluyen las combinaciones de los mismos. Los archivos de datos de usuario pueden contener contenido confidencial o datos confidenciales. Estos datos confidenciales pueden incluir cualquier contenido de usuario que se ajuste a uno o más patrones o esquemas de datos. Los tipos de datos confidenciales ejemplares incluyen números de seguridad social, números de tarjetas de crédito, números de pasaporte, direcciones, números de teléfono u otra información.

Además de la edición o visualización del archivo de datos de usuario, el repartidor de contenido 211 subdivide el contenido de usuario en una o más porciones o “fragmentos” que están en una forma reducida a partir de la forma estructurada o jerárquica original/nativa. El repartidor de contenido 211 puede proporcionar estos fragmentos de contenido al servicio de DLP compartido 121, junto con metadatos de fragmentos para cada fragmento. Los metadatos de fragmento pueden indicar varias propiedades de fragmento, como un intervalo de ubicación del fragmento en el contenido total y una longitud del fragmento. El intervalo de ubicación corresponde a una ubicación del fragmento con relación al documento/archivo de usuario general, y la longitud del fragmento corresponde a un tamaño del fragmento.

El servicio de DLP compartido 121 analiza individualmente los fragmentos de contenido para

identificar los datos confidenciales entre el contenido reducido de usuario de los fragmentos, y proporciona indicaciones de los datos confidenciales al módulo de DLP de aplicación 113. En algunos ejemplos abordados posteriormente, se aplican varios umbrales a conteos o cantidades de datos confidenciales antes de proporcionar indicaciones al módulo de DLP de aplicación 113. Las indicaciones comprenden intervalos para cada uno de los fragmentos que contienen datos confidenciales, longitudes de los fragmentos y opcionalmente indicadores de tipos de datos o esquemas de datos asociados con los datos confidenciales. Las indicaciones de datos confidenciales pueden emplearse para determinar ubicaciones reales o específicas del contenido confidencial entre los datos estructurados del archivo de datos de usuario. Los indicadores de los tipos de datos pueden ser indicadores codificados de manera simbólica o numérica, como valores enteros, que se refieren a una lista de indicadores que el mapeador 213 puede utilizar para identificar los tipos de datos para anotación.

El mapeador 213 puede emplearse para convertir los intervalos y las longitudes en ubicaciones específicas dentro de un documento o archivo de usuario. Los intervalos y las longitudes corresponden a identidades de fragmentos específicos que mantiene el mapeador 213 y se almacenan en asociación con un identificador de sesión. El identificador de sesión puede ser un identificador único que persiste al menos durante la sesión durante la cual el usuario tiene el documento abierto o visualizado. Puede proporcionarse al mapeador 213 los metadatos de fragmentos del repartidor de contenido 211 para formar relaciones asignadas entre los intervalos de fragmentos, las longitudes y los identificadores de sesión. Respondiendo a la recepción de indicaciones de los datos confidenciales, el mapeador 213 puede emplear las relaciones asignadas para identificar ubicaciones aproximadas indicadas para los datos confidenciales dentro de un documento que corresponde al intervalo y longitudes del fragmento. Puesto que los fragmentos pueden abarcar más de un elemento estructural o jerárquico del archivo de datos de usuario, el mapeador 213 puede realizar otros procesos de ubicación para encontrar ubicaciones específicas en el archivo de datos de usuario para los datos confidenciales.

Por ejemplo, los intervalos pueden indicar ubicaciones aproximadas como una fila o columna particular en una hoja de cálculo. Para determinar una ubicación específica, como una celda dentro de la fila o columna indicada, el mapeador 213 puede utilizar los intervalos/longitudes junto con el conocimiento local de los datos estructurados y el archivo de datos de usuario para localizar el contenido confidencial entre los datos estructurados. El mapeador 213 determina de qué parte del archivo de datos de usuario provienen los fragmentos, como filas, columnas, hojas de trabajo asociadas para ejemplos de hojas de cálculo y diapositivas/páginas y objetos asociados para ejemplos de presentación de diapositivas. Otros ejemplos, como los ejemplos de procesamiento de texto, pueden no tener mucha estructura, y el contenido se reduce más fácilmente y los intervalos pueden basarse en el conteo de palabras del documento o un posicionamiento similar.

En algunos ejemplos, las ubicaciones específicas se determinan buscando el contenido confidencial en una ubicación aproximada particular. Cuando múltiples elementos estructurales o elementos jerárquicos están implicados por un intervalo particular, el mapeador 213 puede buscar o recorrer iterativamente cada uno de los elementos para localizar los datos confidenciales. Por ejemplo, si

hay “n” niveles de estructura/jerarquía en un documento, el mapeador 213 puede navegar primero por las jerarquías superiores y después por las inferiores. En los ejemplos de hojas de cálculo, la jerarquía/estructura puede comprender hojas de trabajo con filas y columnas asociadas. En los ejemplos de documentos de presentación de diapositivas, la jerarquía/estructura podría comprender diapositivas/páginas que tengan formas/objetos asociados. Puede recorrerse cada hoja de trabajo y diapositiva indicada por el intervalo para encontrar las celdas u objetos exactos que contienen el contenido confidencial. En otros ejemplos, la ubicación de los datos confidenciales puede realizarse al recrear uno o más fragmentos asociados con la ubicación aproximada y encontrar los datos confidenciales dentro de esos fragmentos recreados para encontrar la ubicación específica de los datos confidenciales.

Una vez que se han determinado las ubicaciones específicas de los datos confidenciales, puede emplearse el anotador 212 para marcar o de otra manera señalar los datos confidenciales a un usuario. Esta anotación puede tomar la forma de un indicador o letrero global que indica al usuario que hay contenido confidencial en el archivo de datos de usuario. Esta anotación puede tomar la forma de indicadores individuales que indican marcas cercanas a los datos confidenciales. En un ejemplo, la figura 2 muestra la configuración 201 con una vista de una interfaz de usuario de hoja de cálculo que tiene un libro actualmente abierto para su visualización o edición. Se muestra una anotación de letrero 220, así como anotaciones de celdas individuales 221. Las anotaciones de celdas individuales 221 comprenden indicaciones gráficas que anotan una o más porciones del contenido de usuario y comprenden indicadores colocados próximos a una o más porciones que pueden seleccionarse en la interfaz de usuario 112 para presentar opciones de ofuscación.

Puede presentarse a un usuario una o más opciones cuando se selecciona una anotación particular. Puede presentarse el menú emergente 202 que incluye varias opciones de visualización/edición, como cortar, copiar, pegar, entre otras. El menú emergente 202 también puede incluir opciones de ofuscación. La selección de una de las opciones de ofuscación puede producir contenido ofuscado que mantiene un esquema de datos del contenido de usuario asociado, y comprende símbolos seleccionados para evitar la identificación del contenido de usuario asociado mientras se mantiene el esquema de datos del contenido de usuario asociado. En algunos ejemplos, los símbolos se seleccionan en parte con base en el esquema de datos del contenido de usuario asociado, entre otras consideraciones. Por ejemplo, si el esquema de datos incluye un esquema de datos numéricos, entonces las letras podrían utilizarse como símbolos de ofuscación. Del mismo modo, si el esquema de datos incluye un esquema de datos alfabéticos, entonces los números podrían utilizarse como símbolos de ofuscación. Las combinaciones de letras y números, u otros símbolos, pueden seleccionarse como símbolos de ofuscación en ejemplos de contenido alfanumérico.

En la figura 2, una primera opción de ofuscación incluye reemplazar el contenido confidencial con texto oculto u ofuscado, mientras una segunda opción de ofuscación incluye reemplazar todo el contenido con un patrón o esquema de datos similar al contenido de la anotación seleccionada actualmente. Por ejemplo, si se incluye un SSN en una celda, se le puede presentar al usuario opciones para reemplazar los dígitos en el SSN con caracteres “X” mientras se deja intacto un esquema de datos del SSN, es decir, se

deja en la disposición conocida de caracteres "3-2-4" separados por caracteres de guion. Además, otra opción de ofuscación puede incluir una opción para reemplazar todos los SSN que se adapten al patrón del SSN seleccionado con caracteres "X". Debe entenderse que pueden presentarse diferentes opciones de ofuscación ejemplar, y pueden utilizarse diferentes caracteres en el proceso de reemplazo. Sin embargo, independientemente de los caracteres de ofuscación empleados, los datos confidenciales se anonimizan, depuran, limpian o se hacen imposibles de identificar como contenido original.

Pasando ahora a la figura 3, se muestra que la configuración 300 ejemplar se enfoca en aspectos del servicio de DLP 121. En la figura 3, el servicio de DLP 121 recibe porciones de contenido reducido de usuario, proporcionadas en uno o más fragmentos de contenido por el repartidor de contenido 211, junto con metadatos de fragmentos que al menos incluyen intervalos en el contenido total y las longitudes de los fragmentos. En la figura 3 se muestran dos tipos de contenido estructurado de usuario, es decir, el contenido de hoja de cálculo 301 y el contenido de presentación de diapositivas/presentación 302. El contenido de hoja de cálculo 301 tiene una estructura que refleja las filas 321 y las columnas 322 que definen celdas individuales. Además, el contenido de la hoja de cálculo 301 puede tener más de una hoja de trabajo 320 que está delimitada por pestañas debajo de la hoja de trabajo, y cada hoja de trabajo puede tener un conjunto separado de filas/columnas. Cada celda puede tener contenido de usuario, como caracteres, contenido alfanumérico, contenido de texto, contenido numérico u otro contenido. El contenido de presentación de diapositivas 302 puede tener una o más diapositivas o páginas 323 que incluyen una pluralidad de objetos 324. Cada objeto puede tener contenido de usuario, como caracteres, contenido alfanumérico, contenido de texto, contenido numérico u otro contenido.

El repartidor de contenido 211 subdivide el contenido de usuario en porciones y elimina cualquier estructura asociada, como al extraer cualquier contenido de usuario, como contenido de texto o alfanumérico, celdas u objetos y después organizar el contenido extraído en fragmentos reducidos o lineales para su entrega al servicio de DLP 121. Estos fragmentos y metadatos de fragmentos se proporcionan al servicio de DLP 121 para el descubrimiento de datos confidenciales potenciales.

Una vez que el servicio de DLP 121 recibe fragmentos individuales del contenido de usuario, el servicio de clasificación 124 realiza diversos procesamiento en los fragmentos. Además, el seguidor 123 mantiene registros de datos 332 que comprenden una o más estructuras de datos que relacionan los intervalos/longitudes y el identificador de sesión para conteos de datos confidenciales encontrados. Los registros de datos 332 se almacenan para ese servicio de DLP 121 para proporcionar los intervalos/longitudes de los fragmentos que contienen datos confidenciales a una aplicación solicitante para ubicar y anotar aún más cualquier contenido confidencial que se encuentre en los mismos.

El servicio de clasificación 124 analiza cada uno de los fragmentos con respecto a varias reglas de clasificación 331 para identificar datos confidenciales o contenido confidencial. Las reglas de clasificación 331 pueden establecer uno o más esquemas de datos predeterminados definidos por una o más expresiones utilizadas para analizar los fragmentos planos/representaciones de datos para identificar partes de los fragmentos como indicativos de uno o más patrones de contenido predeterminados o uno o más tipos de contenido predeterminados.

El contenido confidencial se identifica comúnmente con base en un patrón estructural de datos o "esquema" de datos que está asociado con contenido confidencial. Estos patrones o esquemas pueden identificar cuándo pueden diferirse los contenidos exactos de los fragmentos, pero los datos pueden ajustarse a un patrón o disposición que refleje tipos de datos confidenciales. Por ejemplo, un SSN podría tener una determinada disposición de datos que tenga un número predeterminado de dígitos entremezclados y separados por un número predeterminado de guiones. Las reglas de clasificación 331 pueden incluir varias definiciones y políticas utilizadas en la identificación de datos confidenciales. Estas reglas de clasificación pueden incluir políticas de privacidad, patrones de datos, esquemas de datos y políticas de umbral. Las políticas de privacidad pueden indicar que ciertos datos potencialmente confidenciales pueden no indicarse como sensibles a una aplicación debido a políticas de la empresa, organización o usuario, entre otras consideraciones. Las políticas de umbral pueden establecer umbrales mínimos para encontrar datos confidenciales en los distintos fragmentos antes de que se informe a la aplicación de la presencia de datos confidenciales. Las reglas de clasificación 331 pueden establecerlas los usuarios o los responsables de políticas, como los administradores.

Además, el servicio de clasificación 124 puede procesar el contenido de datos a través de una o más expresiones regulares manejadas por el servicio de expresiones regulares (regex) 333. El servicio de expresiones regulares 333 puede incluir servicios de coincidencia y procesamiento de expresiones regulares, junto con varias expresiones regulares que un usuario o el creador de políticas podría implementar para la identificación de datos confidenciales. Otros ejemplos del servicio de expresiones regulares 333 se analizan posteriormente en la figura 7.

Como ejemplo específico, el proceso de clasificación 341 ilustra varios fragmentos de contenido C1-C8 que son versiones linealizadas de contenido originalmente en una disposición estructural o jerárquica en un documento o archivo de datos de usuario. El servicio de clasificación 124 procesa estos fragmentos para identificar los fragmentos que comprenden datos confidenciales. Si se encuentran datos confidenciales, pueden proporcionarse indicaciones a la aplicación. Las indicaciones pueden comprender intervalos y longitudes para los datos confidenciales, y se proporcionan para el mapeador 213 para ubicar los datos confidenciales dentro de la estructura del archivo de datos de usuario. El servicio de clasificación 124 puede descartar los fragmentos en sí mismos después de procesar cada fragmento para la identificación de datos confidenciales. Dado que los intervalos y longitudes permiten encontrar los datos confidenciales dentro del archivo de datos original, y el contenido original permanece en el archivo de datos (a menos que se hayan producido ediciones intermedias), no es necesario guardar los fragmentos reales una vez procesados.

Para formar los fragmentos, el repartidor de contenido 211 agrupa el contenido alfanumérico, como el texto, en una o más estructuras de datos lineales, como cadenas o BSTR (cadenas básicas o cadenas binarias). El servicio de clasificación 124 procesa las estructuras de datos lineales y determina una lista de resultados. Los fragmentos se comprueban en busca de datos confidenciales, y puede determinarse que porciones de las estructuras de datos lineales tienen contenido confidencial. El servicio de clasificación 124 junto con el seguidor 123 determinan los intervalos/longitudes correspondientes a fragmentos que

contienen datos confidenciales entre las estructuras de datos lineales. Estos intervalos pueden indicar ubicaciones generales que pueden traducirse de nuevo a ubicaciones específicas en el documento original (por ejemplo, archivo de datos de usuario) que contiene el contenido de usuario. Cuando se reciben los fragmentos, el seguidor 123 puede correlacionar cada fragmento con la información de intervalo/longitud indicada en los metadatos del fragmento. Esta información de intervalo/longitud puede utilizarse para asignar de manera inversa a la estructura o jerarquía del documento original mediante el mapeador 213.

Sin embargo, el servicio de DLP 121 comúnmente solo tiene un contexto parcial de regreso al documento original o al archivo de datos de usuario, tal como se indica por los intervalos en las estructuras de datos lineales generadas originalmente. Además, las estructuras de datos lineales y el contenido de usuario podrían liberarse/eliminarse por el servicio de clasificación 124 al final de un proceso de clasificación. Esto puede significar que el servicio de clasificación 124 puede no ser capaz de buscar directamente el contenido confidencial para localizar específicamente el contenido confidencial dentro del documento original, e incluso si el servicio de clasificación 124 pudiera buscar el contenido confidencial preciso el servicio de clasificación 124 podría no ser capaz de encontrar el contenido confidencial porque el algoritmo de "fragmentación" podría cruzar los límites de las construcciones o estructuras jerárquicas en el documento o archivo de datos original. Como ejemplo específico, la hoja de trabajo 320 en un documento de hoja de cálculo puede tener el texto "SSN 123 45 6789" que abarca cuatro celdas adyacentes. Ventajosamente, el servicio de clasificación 124 puede encontrar este texto como contenido confidencial. Sin embargo, debido al análisis de cruce de límites por el servicio de clasificación 124, al final de la evaluación de la regla de política, el servicio de clasificación 124 generalmente no tiene suficientes datos para encontrar el contenido confidencial en el documento original para presentarlo a un usuario. Un usuario puede quedarse con una impresión incorrecta de que no existe contenido confidencial presente.

Para escanear eficientemente el contenido de usuario en busca de contenido confidencial, el servicio de clasificación 124 lee un fragmento de contenido de usuario en un momento durante la inactividad de la aplicación, realiza un análisis parcial y continúa el proceso. Cuando el servicio de clasificación 124 termina de leer todo el contenido, el servicio de clasificación 124 solo tiene posiciones aproximadas para el contenido confidencial en el contenido original, solo como un inicio/intervalo y una longitud. Para volver a asignar un documento estructurado o semiestructurado de manera eficiente, el mapeador 213 puede emplear una combinación de técnicas. Cabe señalar que estas técnicas difieren de cómo podría funcionar un corrector ortográfico o gramatical, en parte debido a que puede requerirse el contenido total, en lugar de solo una palabra/oración/párrafo, para entender si el contenido excedió un umbral.

Para cada nivel de jerarquía o estructura física presente en el documento original (es decir, hojas de trabajo en un libro o diapositivas en una presentación), el mapeador 213 utiliza un identificador para indicar la existencia en una estructura de datos de asignación, y además subdividir por un número razonable de niveles de jerarquía (es decir, filas en una hoja de trabajo, formas en una diapositiva) el contenido de manera que a medida que se procesa cada uno, el mapeador 213 realice un seguimiento de la longitud del contenido original y, de acuerdo con el orden de inserción en la asignación, el inicio implícito de ese elemento. El identificador puede ser un identificador vigente durante el proceso que persiste entre

instancias abiertas de un documento particular, o puede ser diferente en cada instancia del documento particular. En algunos ejemplos, los cálculos para combinar la presencia/ausencia de contenido confidencial se retienen hasta que no quede contenido sin procesar ni ediciones pendientes que modifiquen aún más el contenido.

Suponiendo que existe contenido confidencial, el mapeador 213 recibe del servicio de DLP 121 un inicio y una longitud de cada pieza de contenido confidencial y el mapeador 213 realiza una búsqueda en la estructura de datos de asignación de los identificadores e inserciones del contenido confidencial dentro de la región mapeada más precisa para encontrar la ubicación exacta. Por razones de rendimiento, solo puede seguirse un cierto número de niveles de jerarquía, de modo que una tabla dentro de una forma dentro de una diapositiva, o una celda dentro de una fila dentro de una hoja de trabajo no se pueda seguir individualmente. Por lo tanto, puede realizarse un recorrido parcial después de realizar un mapeo inverso para encontrar la ubicación precisa.

En un ejemplo específico, un libro podría tener 20 hojas de trabajo, pero millones de filas y cada uno de los millones de filas podrían tener 50 columnas de datos de usuario. Para un número relativamente pequeño de datos confidenciales en estos (es decir, una hoja tiene solo una columna con datos confidenciales), el proceso de clasificación puede consumir demasiada memoria para tener $20 * 1 \text{ millón} * 50$ piezas recordadas de "longitud + intervalo" de datos. Eliminar la última dimensión supone un ahorro de 50x en memoria, por un pequeño costo de cómputo en el momento en el cual los datos confidenciales se identifican realmente en el documento original. Ventajosamente, puede mantenerse un pequeño espacio ocupado de memoria para el mapeo inverso del inicio/longitudes nuevamente al contenido original.

Para ilustrar adicionalmente la operación de los elementos de las figuras 1-3, se presenta un diagrama de flujo en la figura 4. Dos flujos principales se presentan en la figura 4, es decir, un primer flujo 400 para la identificación de datos confidenciales, y un segundo flujo 401 para anotaciones y ofuscación de datos confidenciales. El primer flujo 400 puede alimentar al segundo flujo 401, aunque son posibles otras configuraciones.

En la figura 4, el servicio de DLP 121 recibe (410) subconjuntos de contenido estructurado de usuario consolidado en representaciones reducidas asociadas, cada una de las representaciones reducidas asociadas tiene una correspondencia con un subconjunto correspondiente del contenido estructurado de usuario. Tal como se mencionó anteriormente, el contenido estructurado puede comprender contenido de hoja de cálculo organizado en hojas/filas/columnas, o puede incluir otras estructuras como contenido de presentación de diapositivas organizado en diapositivas/objetos, contenido de programa de dibujo organizado en páginas/objetos, o contenido de texto organizado en páginas, entre otras estructuras. Estos subconjuntos del contenido estructurado de usuario pueden incluir los "fragmentos" 141-146 mostrados en la figura 1 o los fragmentos C₁-C₈ en la figura 3, entre otros. La estructura del contenido subyacente de usuario se reduce o elimina en estos subconjuntos para formar los fragmentos, y cada uno de los subconjuntos puede volver a la estructura original haciendo referencia a identificadores o localizadores estructurales, como, por ejemplo, hojas/filas/columnas o diapositivas/objetos.

El servicio de DLP 121 recibe estos fragmentos y metadatos de fragmentos, como a través del

enlace 160 o API 122 en la figura 1, y analiza individualmente (411) las representaciones reducidas para clasificar porciones que comprenden contenido confidencial correspondiente a uno o más esquemas de datos predeterminados. Las reglas de clasificación 125 pueden establecer uno o más esquemas de datos predeterminados definidos por una o más expresiones utilizadas para analizar los fragmentos/representaciones de datos reducidos para identificar porciones de los fragmentos como indicativos de uno o más patrones de contenido predeterminados o uno o más tipos de contenido predeterminados.

Si se encuentran datos confidenciales (412), entonces, para cada una de las porciones, el servicio de DLP 121 determina (413) un intervalo/longitud asociado con relación al contenido estructurado de usuario indicado como mantenido en el seguidor 123 en los registros de datos 332. El servicio DLP 121 entonces indica (414) al menos el intervalo/longitud asociado para las porciones a la aplicación de usuario 111 para marcar el contenido confidencial en la interfaz de usuario 112 para la aplicación de usuario 111. Si no se encuentran datos confidenciales, o si no se alcanzan los umbrales asociados, entonces el procesamiento posterior de los fragmentos puede continuar o la monitorización adicional de fragmentos adicionales según lo dispuesto por la aplicación de usuario 111. Además, la edición o el cambio del contenido de usuario puede generar procesos de clasificación adicionales o repetidos para cualquier contenido de usuario modificado o editado.

El módulo de DLP de aplicación 113 recibe (415) desde el servicio de clasificación del servicio de DLP 121 indicaciones de una o más porciones del contenido de usuario que incluyen el contenido confidencial, donde las indicaciones comprenden intervalos/longitudes asociadas con el contenido confidencial. El módulo de DLP de aplicación 113 presenta (416) indicaciones gráficas en la interfaz de usuario 112 a la aplicación de usuario 111 que anotan una o más porciones del contenido de usuario como contenido confidencial. El módulo de DLP de aplicación 113 puede presentar (417) opciones de ofuscación en la interfaz de usuario 112 para ocultar el contenido confidencial dentro de al menos una porción seleccionada entre una o más porciones del contenido de usuario. En respuesta a una selección de usuario de al menos una de las opciones de ofuscación, el módulo de DLP de aplicación 113 reemplaza (418) el contenido de usuario asociado con el contenido ofuscado que mantiene un esquema de datos del contenido de usuario asociado.

La figura 5 ilustra el diagrama de secuencia 500 para ilustrar adicionalmente el funcionamiento de los elementos de las figuras 1-3. Además, la figura 5 incluye un ejemplo detallado de la estructura 510 para algunas de las etapas del proceso en la figura 5. En la figura 5, la aplicación 111 podría abrir un documento para que un usuario lo visualice o edite. Este documento puede detectarse por el módulo de DLP de aplicación 113. Cualquier política asociada o regla de clasificación puede enviarse al servicio de DLP 121 para definir cualquier política de clasificación. El servicio 121 de DLP puede mantener una instancia de procesamiento del documento abierto en el registro 332, que podría incluir una lista de varios documentos abiertos. Cuando el módulo DLP 113 detecta los períodos de tiempo de procesamiento inactivo de la aplicación 111, puede presentarse un indicador inactivo al servicio de DLP 121, que en respuesta solicita fragmentos de contenido de usuario para su clasificación. Alternativamente, el módulo DLP 113 puede

enviar fragmentos de contenido de usuario al servicio de DLP 121 durante los períodos de inactividad de la aplicación 111. El módulo DLP 113 divide el contenido de usuario en fragmentos, y estos fragmentos pueden determinarse con base en el texto u otro contenido incluido en estructuras u objetos jerárquicos del documento. Una vez que se determinaron los fragmentos, el módulo de DLP 113 transfiere los fragmentos al servicio de DLP 121 para su clasificación. El servicio de DLP 121 clasifica cada fragmento individualmente y aplica reglas de clasificación a los fragmentos para identificar el contenido de usuario potencialmente confidencial entre los fragmentos. Este proceso de clasificación puede ser un proceso iterativo para garantizar que se procesen todos los fragmentos transferidos por el módulo de DLP 113. Si se encuentran datos o contenido confidencial entre los fragmentos, el servicio de DLP 121 indica la presencia de datos confidenciales en el módulo de DLP 113 para su posterior manejo. Tal como se menciona en el presente documento, los datos confidenciales pueden indicarse mediante intervalos, ubicaciones aproximadas u otra información de ubicación, así como información de longitud. El módulo de DLP 113 puede realizar uno o más procesos de anotación y procesos de ofuscación en los datos confidenciales del documento.

Las reglas de clasificación pueden establecerse antes del proceso de clasificación, como por parte de los usuarios, administradores, personal de políticas u otras entidades. Tal como se observa en la estructura 510, varias reglas 511 y 512 pueden basarse en uno o más predicados. Los predicados se muestran en dos categorías en la figura 5, los predicados relacionados con contenido 511 y predicados relacionados con acceso 512. Los predicados relacionados con contenido 511 pueden comprender esquemas de datos que indican datos confidenciales, como patrones de datos, información estructural de datos o expresiones regulares que definen los esquemas de datos. Los predicados relacionados con el acceso 512 comprenden el nivel de usuario, el nivel de organización u otras reglas basadas en el acceso, como las reglas de uso compartido de contenido que definen cuándo no se desea que los datos confidenciales sean difundidos o liberados por parte de usuarios, organizaciones u otros factores particulares.

Pueden establecerse reglas de política 513 que combinan uno o más predicados relacionados con contenido y predicados relacionados con acceso en las políticas 551-554. Cada regla de política también tiene una prioridad y una acción asociada. En general, la prioridad coincide con la rigurosidad de la acción. Por ejemplo, una regla de política podría definir que las funciones de “guardar” de la aplicación deben bloquearse. En otra regla de política ejemplar, el contenido de usuario puede contener los SSN que se definen de acuerdo con un predicado relacionado con contenido, pero de acuerdo con un predicado relacionado con acceso, estos SSN pueden ser aceptables para su difusión. La mayoría de las reglas de política contienen al menos un predicado de clasificación entre los predicados 511-512. Estas políticas pueden afectar una o más acciones 514. Las acciones pueden incluir varias operaciones de anotación que una aplicación podría realizar en respuesta a la identificación o al contenido confidencial, como la notificación de un usuario, notificación pero que permite la anulación de un usuario, bloqueo de características/funciones (es decir, funciones “guardar” o “copiar”), y anulaciones justificadas, entre otras.

La figura 6 ilustra el diagrama de flujo 600 para ilustrar adicionalmente la operación de los

elementos de las figuras 1-3. La figura 6 se centra en un proceso completo ejemplar de identificación de datos confidenciales, anotaciones y procesos de ofuscación. El subproceso 601 comprende el establecimiento, almacenamiento y recuperación de políticas y reglas. Estas políticas y reglas pueden incluir reglas de anotación, reglas de clasificación, expresiones regulares, políticas organizacionales/de usuario, entre otra información abordada en el presente documento. En la operación 611 de la figura 6, pueden introducirse varias reglas de detección 630 y reglas de reemplazo 631 a través de una interfaz de usuario o API para configurar políticas de detección. Las reglas de detección 630 y las reglas de reemplazo 631 pueden comprender varios predicados y reglas tal como se encuentran en la figura 5, entre otros. Los usuarios, administradores, personal de políticas u otras entidades pueden introducir las reglas de detección 630 y las reglas de reemplazo 631, como establecer políticas para usuarios, organizaciones o el uso de aplicaciones, entre otras entidades y actividades. Las reglas de detección 630 y las reglas de reemplazo 631 pueden almacenarse en uno o más sistemas de almacenamiento en operación 612 para su uso posterior. Cuando uno o más clientes desean utilizar las políticas establecidas por las reglas de detección 630 y las reglas de reemplazo 631, estas políticas pueden descargarse o recuperarse en la operación 613. Por ejemplo, una aplicación puede descargar las reglas de anotación para anotar el contenido confidencial en una interfaz de usuario, mientras las reglas de clasificación pueden descargarse por un servicio de DLP compartido para clasificar el contenido de usuario como contenido confidencial.

El subproceso 602 comprende actividades de aplicación cliente, como cargar documentos para editar o visualizar en una interfaz de usuario, y proporcionar fragmentos de esos documentos para su clasificación. En la operación 614, una aplicación cliente puede proporcionar una o más experiencias de usuario final para procesar contenido de usuario, editar contenido de usuario o visualizar contenido de usuario, entre otras operaciones. La operación 614 también puede proporcionar procesos de anotación y ofuscación que se analizan más adelante. La operación 615 proporciona porciones de este contenido de usuario a un servicio de DLP compartido para clasificar el contenido de usuario. En algunos ejemplos, las porciones comprenden fragmentos reducidos de contenido de usuario que se despojan de la estructura o jerarquía asociada del documento original.

El subproceso 603 comprende la clasificación del contenido de usuario para detectar datos confidenciales entre el contenido de usuario, así como la anotación de estos datos confidenciales para un usuario. En la operación 616, se aplican varias reglas de detección, como las expresiones regulares que se analizan a continuación en la figura 7, entre otras reglas y procesos de detección. Si se encuentran datos confidenciales, la operación 617 determina si se debe notificar a un usuario. Es posible que la notificación no ocurra si la cantidad de datos confidenciales disminuye por debajo de un umbral de alerta. Sin embargo, si se debe alertar al usuario, la operación 619 puede calcular las ubicaciones de los datos confidenciales dentro de las regiones detectadas de los datos estructurados. Tal como se aborda en el presente documento, puede emplearse un proceso de asignación para determinar ubicaciones específicas de datos confidenciales dentro de elementos estructurados o elementos jerárquicos a partir de intervalos de datos reducidos y longitudes de las cadenas o porciones de datos confidenciales. Una vez que se determinan estas ubicaciones específicas, la operación 618 puede mostrar las ubicaciones al usuario. Las

anotaciones u otros elementos destacados de la interfaz de usuario se emplean para indicar al usuario que existen datos confidenciales entre el contenido de usuario.

El subproceso 604 comprende la ofuscación de datos confidenciales dentro del contenido de usuario que comprende los elementos estructurados o jerárquicos. En la operación 621, puede recibirse la entrada de usuario para reemplazar al menos una instancia de datos confidenciales con datos/texto “seguros” u ofuscados. Cuando a un usuario se le muestra una región resaltada que muestra un fragmento de datos confidenciales que provocó la aparición de una anotación o una “sugerencia de política”, se le puede presentar al usuario una opción para reemplazar los datos confidenciales con “texto seguro” que ofusca los datos confidenciales. Dependiendo de las elecciones realizadas por las entidades que inicialmente establecen las políticas en la operación 611, las operaciones 622 y 624 determinan y generan una o más reglas de reemplazo u ofuscación. Las reglas de ofuscación pueden utilizarse para reemplazar un nombre clave interno con un nombre aprobado comercialmente, para ofuscar información de identificación personal (PII, por sus siglas en inglés) con nombres estándar, pueden utilizarse para reemplazar datos numéricos confidenciales con un conjunto de caracteres que indican a los futuros lectores del documento sobre el tipo de datos confidenciales (es decir, números de tarjetas de crédito, números de seguro social, números de identificación de vehículos, entre otros) sin revelar los datos confidenciales reales. La operación 623 reemplaza los datos confidenciales con los datos ofuscados. Los datos ofuscados pueden utilizarse para reemplazar datos confidenciales numéricos con un conjunto de caracteres que podrían utilizarse para confirmar un esquema de datos o tipo de contenido, pero siguen siendo insuficientes para obtener los datos originales incluso por parte de un individuo determinado (es decir, para determinar que la pieza de contenido es un SSN pero no revelar el SSN real). Los usuarios pueden realizar el reemplazo individual o de única instancia de contenido confidencial con texto ofuscado, o el reemplazo masivo desde una interfaz de usuario que muestra múltiples instancias de contenido confidencial.

El reemplazo de contenido confidencial, como texto o contenido alfanumérico, puede realizarse con expresiones regulares, o alternativamente a través de autómatas finitos no deterministas (NFA, por sus siglas en inglés), autómatas finitos deterministas (DFA, por sus siglas en inglés), autómatas de pila (PDA), máquinas de Turing, código funcional arbitrario u otros procesos. El reemplazo de contenido confidencial generalmente comprende la coincidencia de patrones entre texto o contenido. Esta coincidencia de patrones puede dejar caracteres o contenido sin ocultar considerando si el patrón de destino tiene la capacidad de que existan múltiples caracteres en una ubicación específica en una cadena y esos caracteres no necesitan ocultarse, como los caracteres delimitadores. Por ejemplo, la cadena "123-12-1234" podría convertirse en "xxx-xx-xxxx" y la cadena "123 12 1234" podría convertirse en "xxx xx xxxx" después de un proceso de ocultación. Esta coincidencia de patrones también puede mantener ciertas porciones discernibles para propósitos de unicidad, como con el último número predeterminado de dígitos de un número de tarjeta de crédito o SSN. Por ejemplo, "1234-1234-1234-1234" podría convertirse en "xxxx-xxxx-xxxx-1234" después de un proceso de ocultación. Para la ocultación/reemplazo de nombres clave, no todos los aspectos son patrones y pueden ser nombres clave internos u otras palabras clave. Por ejemplo, un nombre clave "Whistler" podría convertirse en "Windows XP" después de un proceso de

ocultación. Además, puede permitirse que los patrones que reemplazan un número variable de caracteres con texto seguro mantengan una longitud consistente o establezcan la longitud en una constante conocida. Por ejemplo, la misma regla puede convertir "1234-1234-1234-1234" en "xxxx-xxxx-xxxx-1234" y "xxxxx-xxxxx-x1234" después de un proceso de ocultación. Esto puede requerir un patrón que contenga datos suficientes para manejar cualquiera de estos casos. Las expresiones regulares pueden manejar tales casos aumentando la expresión regular rodeando cada expresión de coincidencia de átomos con paréntesis, y haciendo un seguimiento de qué declaraciones de "coincidencia" aumentadas se emparejan con qué declaraciones de "reemplazo". Otros ejemplos de coincidencia de expresiones regulares se observan posteriormente en la figura 7.

Para mantener la integridad de los procesos de anotación y clasificación entre más de un documento/archivo, pueden establecerse varios procesos. Las reglas y políticas de detección/clasificación, anotación y ofuscación generalmente no se incluyen en los archivos de documentos. Esto permite cambios en las políticas y evita la ingeniería inversa de las técnicas de ofuscación. Por ejemplo, si un usuario guarda un documento, cierra y carga el mismo documento, entonces pueden haber cambiado las reglas sobre qué porciones del documento contienen los datos confidenciales necesarios para considerar la presencia de datos confidenciales una incidencia de política. Además, los indicadores de anotación no deben incluirse en las operaciones del portapapeles, como cortar, copiar o pegar. Si un usuario copia contenido de un documento y lo pega en otro, ese segundo documento podría tener diferentes reglas de detección/clasificación, anotación y ofuscación aplicadas. Si un usuario lleva el texto contenido de un primer documento y lo pega en un segundo documento, las anotaciones del primer documento deberían considerarse irrelevantes hasta que se vuelvan a clasificar. Incluso, si un usuario copia contenido de un documento en el mismo documento, cualquier conteo del contenido confidencial podría cambiar y aquello que debe resaltarse en todo el documento podría cambiar.

La figura 7 ilustra el diagrama de flujo 700 para ilustrar adicionalmente la operación de los elementos de las figuras 1-3. La figura 7 se centra en las operaciones de expresión regular durante los procesos de ofuscación de datos confidenciales. En la figura 7, dada una expresión regular (regex), como la expresión regular 730 del ejemplo de licencia de conducir ficticia, y una cadena que coincide con ella, puede generarse una coincidencia completa al menos al aumentar la expresión regular al rodear cada expresión de coincidencia de caracteres separables con paréntesis (por ejemplo, cada átomo), tal como se indica en la operación 711. La expresión regular aumentada puede volver a aplicarse o ejecutarse en la operación 712 para realizar un proceso de ofuscación u ocultación. Para cada coincidencia, las operaciones 713-714 determinan los conjuntos de caracteres más amplios y estrechos que realmente coinciden. Por ejemplo, cuando el carácter coincidente es "-", el carácter es estrecho ya que es un solo carácter. Cuando el carácter coincidente es el conjunto de todos los caracteres alfabéticos, es amplio. El conteo absoluto de caracteres que podría estar en cualquier región es el determinante esencial. Un proceso de ofuscación en la operación 715 puede reemplazar caracteres de acuerdo con una amplitud de coincidencia. Para aquellos caracteres coincidentes que son caracteres individuales, un proceso de ofuscación no puede realizar ningún cambio. Para aquellos caracteres coincidentes que se encuentran en grupos amplios, un proceso

de ofuscación reemplaza los caracteres con un carácter "seguro" que no es miembro del conjunto. Por ejemplo, un conjunto de todas las letras se convierte en "0", un conjunto de todos los números se convierte en "X" y el contenido alfanumérico mixto se convierte en "?", con una lista alternativa de caracteres que se utilizarán hasta que se agoten. Una vez que el texto o contenido pasó por un proceso de ofuscación u ocultación, la operación 716 confirma que el texto o contenido se ofuscó con éxito cuando la nueva cadena de texto/contenido ya no coincide con la expresión regular original.

La figura 8 ilustra el diagrama gráfico 800 para ilustrar adicionalmente la operación de los elementos de las figuras 1-3. La figura 8 se centra en los procesos de umbral mejorados utilizados en la anotación de datos confidenciales en las interfaces de usuario. Las operaciones de la figura 8 pueden comprender operaciones de histéresis mejoradas para anotar datos confidenciales, y los administradores de políticas o usuarios, entre otras entidades, pueden establecer varios umbrales o reglas de anotación.

La figura 8 incluye la gráfica 800 que incluye un eje vertical que indica una cantidad de datos/elementos de contenido confidenciales presentes en un documento, y un eje horizontal que indica el tiempo. Se establece un primer umbral 820 que puede iniciar la presentación o eliminación de las anotaciones de contenido confidencial en una interfaz de usuario. Puede establecerse un segundo umbral 822 que también puede iniciar la presentación o eliminación de las anotaciones de contenido confidencial. Puede establecerse un factor de elasticidad 821 y la propiedad de resiliencia 823 para modificar el comportamiento del primer y segundo umbral.

Cuando se han anotado datos confidenciales en una interfaz de usuario, como mediante indicadores, marcas o resaltados, un usuario puede editar el contenido confidencial para solucionar incidencias de contenido confidencial (como al seleccionar una o más opciones de ofuscación). Sin embargo, una vez que se resolvió un número umbral de incidencias de contenido confidencial, es posible que no existan suficientes instancias restantes de una incidencia para garantizar la anotación del documento como una infracción general de las reglas de contenido confidencial para la organización o ubicación de guardado. Del mismo modo, cuando se introduce nuevo contenido confidencial en un documento, pueden existir instancias suficientes para garantizar la anotación del documento para indicar el contenido confidencial a un usuario.

Durante los procesos de edición de contenido por parte de los usuarios, la habilitación e inhabilitación de indicadores de anotación para uno o más elementos de contenido puede basarse, al menos en parte, en una cantidad actual de los elementos de contenido con respecto a las reglas de anotación. Las reglas de anotación pueden comprender al menos la primera cantidad umbral 820, el factor de elasticidad 821 para modificar la primera cantidad umbral 820 a una segunda cantidad umbral 822 cuando está habilitado, y una indicación de una resiliencia umbral o propiedad de "cohesión" 823 que indica cuándo la segunda cantidad umbral 822 anula la primera cantidad umbral 820. Un servicio de anotación, como el anotador 212, puede determinar o identificar las reglas de anotación como las reglas de política 513 y las acciones 514 abordadas en la figura 5 que se establecen para las entidades objetivo asociadas con la edición de contenido. Las entidades objetivo pueden incluir usuarios que realizan la edición de contenido, una organización que comprende al usuario que realiza la edición de contenido o un tipo de

aplicación de la aplicación de usuario, entre otras. Durante la edición de usuario de un documento que incluye contenido confidencial o que podría incluir contenido confidencial, el anotador 212 monitoriza el contenido de usuario en un archivo de datos de usuario asociado presentado para la edición de contenido en una interfaz de usuario para la aplicación de usuario. El anotador 212 identifica una cantidad de elementos de contenido que incluyen contenido confidencial entre el contenido de usuario correspondiente a uno o más esquemas de datos predeterminados abordados en el presente documento. Los elementos de contenido pueden incluir celdas, objetos, formas, palabras u otros elementos de datos estructurales o elementos jerárquicos de datos.

Durante la edición, y con base al menos en la cantidad de elementos de contenido que excede una primera cantidad umbral, el anotador 212 inicia la presentación de al menos un indicador de anotación en la interfaz de usuario que indica el contenido de usuario en la interfaz de usuario que contiene al menos el primer contenido confidencial. En la figura 8 (comenzando con las anotaciones en un estado “inactivo”), el primer umbral 820 indica una cantidad ejemplar de “8” en el punto de transición 830 como activación de la presentación de indicadores de anotación en una interfaz de usuario. La cantidad de elementos de contenido con contenido confidencial puede aumentar, por ejemplo, mediante la edición de usuario, y entonces puede disminuir después de que un usuario observe que existe contenido confidencial y comience a seleccionar opciones de ofuscación para ocultar este contenido confidencial.

Con base al menos en la cantidad de elementos de contenido que inicialmente exceden la primera cantidad umbral 820 y que posteriormente se encuentran por debajo de la primera cantidad umbral 820 cuando el factor de elasticidad 821 se aplica a la primera cantidad umbral 820, el anotador 212 establece la segunda cantidad umbral 822 con base al menos en el factor de elasticidad. Cuando la segunda cantidad umbral 822 está activa (es decir, cuando el factor de elasticidad 821 se aplica a la primera cantidad umbral 820), la segunda cantidad umbral 822 se utiliza para iniciar la eliminación de la presentación de al menos un indicador de anotación cuando la cantidad se encuentra por debajo de la segunda cantidad umbral 822, tal como se observa en el punto de transición 832. Sin embargo, con base al menos en la cantidad de elementos de contenido que inicialmente excede la primera cantidad umbral 820 y después se encuentra por debajo de la primera cantidad umbral 820 cuando el factor de elasticidad no se aplica a la primera cantidad umbral 820, se elimina la presentación de al menos un indicador de anotación, tal como lo indica el punto de transición 831.

El factor de elasticidad 821 puede comprender un porcentaje que oscila entre 0 y 100 por ciento, u otra métrica. En un ejemplo específico, se podría establecer una regla de anotación que defina la inclusión de más de 100 SSN en un documento que infringe la política corporativa. Durante la edición de un documento que excede los 100 SSN, una regla de anotación para un primer umbral de cantidad podría hacer resaltar todos los SSN en el documento. Cuando un usuario comienza ofuscando los SSN, se reducirá la cantidad de SSN no ofuscados restantes. El factor de elasticidad puede mantener la anotación o el resaltado de los SSN incluso si ya no se cumple la primera cantidad umbral 820 que activó la anotación, como cuando 99 SSN permanecen sin ofuscar. Un factor de elasticidad de 100 correspondería a una primera cantidad umbral no modificada, y una elasticidad de 0 correspondería a las anotaciones que nunca

se eliminarán hasta que se ofusquen todos los SSN. Un valor intermedio de 50 para el factor de elasticidad correspondería a la eliminación de las anotaciones una vez que se fija el elemento 50 después de que las anotaciones se activen inicialmente para su presentación. Así, en el ejemplo de la figura 8, el factor de elasticidad establece una segunda cantidad umbral para la eliminación de las anotaciones una vez que las anotaciones se presenten a un usuario. En este ejemplo, la segunda cantidad umbral 822 está en “2” y, por lo tanto, cuando las incidencias de contenido confidencial restantes se encuentran por debajo de “2” restantes, las anotaciones se eliminarán, tal como se indica en el punto de transición 832.

Si la segunda cantidad umbral 822 se redujo, y después incidencias adicionales de contenido confidencial aumentaron durante la edición de contenido, entonces el anotador 212 debe decidir cuándo alertar al usuario al presentar las anotaciones nuevamente. Con base al menos en la cantidad de elementos de contenido que inicialmente se encuentran por debajo de la segunda cantidad umbral 822 y después exceden la segunda cantidad umbral 822 cuando la propiedad de resistencia umbral 823 se aplica a la segunda cantidad umbral 822, el anotador 212 inicia la presentación de otras anotaciones en la interfaz de usuario que marca el contenido de usuario en la interfaz de usuario que contiene contenido confidencial, tal como se indica en el punto de transición 833.

La propiedad de resiliencia 823 comprende una propiedad de "cohesión" para la segunda cantidad umbral 822, y se define por una condición de activación/desactivación o booleana. Cuando está deshabilitada, la segunda cantidad umbral 822 no se utiliza para volver a presentar las anotaciones si se supera. Cuando está habilitado, la segunda cantidad umbral 822 se utiliza para volver a presentar las anotaciones si se supera. Por lo tanto, con base al menos en la cantidad de elementos de contenido que inicialmente se encuentran por debajo de la segunda cantidad umbral 822 y después excede la segunda cantidad umbral 822 cuando la propiedad de resiliencia no se aplica a la segunda cantidad umbral 822, el anotador 212 retiene la presentación de las anotaciones que marcan el contenido de usuario en la interfaz de usuario que incluye al menos el contenido confidencial hasta que la cantidad de elementos de contenido exceda nuevamente la cantidad umbral 820.

Pasando ahora a la figura 9, se presenta el sistema de cómputo 901. El sistema de cómputo 901 que es representativo de cualquier sistema o conjunto de sistemas donde pueden implementarse diversas arquitecturas, casos y procesos operativos descritos en el presente documento. Por ejemplo, el sistema de cómputo 901 puede utilizarse para implementar cualquiera de la plataforma de usuario 110 o de la plataforma de DLP 120 de la figura 1. Los ejemplos del sistema de cómputo 901 incluyen, entre otros, servidores, sistemas de computación en la nube, sistemas de cómputo distribuidos, sistemas de red definidos por *software*, computadoras, computadoras de escritorio, computadoras híbridas, servidores en bastidor, servidores web, plataformas de computación en la nube y equipos de centros de datos, así como cualquier otro tipo de máquina de tipo servidor físico o virtual, y otros dispositivos y sistemas de cómputo, así como cualquier variación o combinación de los mismos. Cuando porciones del sistema de cómputo 901 se implementan en dispositivos de usuario, los dispositivos ejemplares incluyen teléfonos inteligentes, computadoras portátiles, tabletas, computadoras de escritorio, sistemas de juegos, sistemas de entretenimiento y similares.

El sistema de cómputo 901 puede implementarse como un solo aparato, sistema o dispositivo o puede implementarse de manera distribuida como múltiples aparatos, sistemas o dispositivos. El sistema de cómputo 901 incluye, entre otros, el sistema de procesamiento 902, el sistema de almacenamiento 903, el *software* 905, el sistema de interfaz de comunicación 907 y el sistema de interfaz de usuario 908. El sistema de procesamiento 902 está operativamente conectado con el sistema de almacenamiento 903, el sistema de interfaz de comunicación 907 y el sistema de interfaz de usuario 908.

El sistema de procesamiento 902 carga y ejecuta el *software* 905 desde el sistema de almacenamiento 903. El *software* 905 incluye el entorno de DLP de aplicación 906 y/o el entorno de DLP compartido 909, el cual es representativo de los procesos abordados con respecto a las figuras anteriores. Cuando se ejecuta mediante el sistema de procesamiento 902 para procesar el contenido de usuario para la identificación, anotación y ofuscación de contenido confidencial, el *software* 905 dirige el sistema de procesamiento 902 para que opere tal como se describe en el presente documento para al menos los diversos procesos, casos operativos y entornos abordados en las implementaciones anteriores. El sistema de cómputo 901 puede incluir opcionalmente dispositivos, prestaciones o funcionalidad adicionales no abordados por fines de brevedad.

Aún con referencia a la figura 9, el sistema de procesamiento 902 puede comprender un microprocesador y circuito de procesamiento que recupera y ejecuta el *software* 905 del sistema de almacenamiento 903. El sistema de procesamiento 902 puede implementarse dentro de un único dispositivo de procesamiento, pero también puede distribuirse entre múltiples dispositivos de procesamiento o subsistemas que cooperan en la ejecución de las instrucciones de programa. Los ejemplos del sistema de procesamiento 902 incluyen unidades de procesamiento central de propósito general, procesadores específicos de aplicación y dispositivos lógicos, así como cualquier otro tipo de dispositivo de procesamiento, combinaciones o variaciones de los mismos.

El sistema de almacenamiento 903 puede comprender cualquier medio de almacenamiento legible por computadora legible por el sistema de procesamiento 902 y capaz de almacenar el *software* 905. El sistema de almacenamiento 903 puede incluir medios volátiles y no volátiles, extraíbles y no extraíbles implementados en cualquier método o tecnología para el almacenamiento de información, como instrucciones legibles por computadora, estructuras de datos, módulos de programa u otros datos. Los ejemplos de medios de almacenamiento incluyen memoria de acceso aleatorio, memoria de solo lectura, discos magnéticos, memoria resistiva, discos ópticos, memoria flash, memoria virtual y memoria no virtual, casetes magnéticos, cinta magnética, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio de almacenamiento adecuado. En ningún caso el medio de almacenamiento legible por computadora es una señal propagada.

Además de los medios de almacenamiento legibles por computadora, en algunas implementaciones, el sistema de almacenamiento 903 también puede incluir medios de comunicación legibles por computadora sobre los cuales al menos parte del *software* 905 puede comunicarse de manera interna o externa. El sistema de almacenamiento 903 puede implementarse como un único dispositivo de almacenamiento, pero también puede implementarse en múltiples dispositivos de almacenamiento o

subsistemas ubicados o distribuidos uno con respecto al otro. El sistema de almacenamiento 903 puede comprender elementos adicionales, como un controlador, capaz de comunicarse con el sistema de procesamiento 902 o posiblemente con otros sistemas.

El *software* 905 puede implementarse en instrucciones de programa y, entre otras funciones, puede, cuando se ejecute mediante el sistema de procesamiento 902, el sistema de procesamiento directo 902 para operar tal como se describe con respecto a los diversos casos operativos, secuencias y procesos ilustrados en el presente documento. Por ejemplo, el *software* 905 puede incluir instrucciones de programa para implementar los entornos de procesamiento de conjuntos de datos y las plataformas abordadas en el presente documento.

En particular, las instrucciones de programa pueden incluir varios componentes o módulos que cooperan o interactúan para llevar a cabo los diversos procesos y casos operativos descritos en el presente documento. Los diversos componentes o módulos pueden incorporarse en instrucciones compiladas o interpretadas, o en alguna otra variación o combinación de instrucciones. Los diversos componentes o módulos pueden ejecutarse de manera síncrona o asíncrona, en serie o simultáneamente, en un entorno de subproceso único o de múltiples subprocesos, o de acuerdo con cualquier otro paradigma de ejecución, variación o combinación de los mismos. El *software* 905 puede incluir procesos, programas o componentes adicionales, como el *software* del sistema operativo (OS, por sus siglas en inglés) u otro *software* de aplicación, además del entorno de DLP de aplicación 906 o del entorno de DLP compartido 909. El *software* 905 también puede comprender *firmware* u otra forma de instrucciones de procesamiento legibles por máquina ejecutables por el sistema de procesamiento 902.

En general, el *software* 905 puede, cuando se carga en el sistema de procesamiento 902 y se ejecuta, transformar un aparato, sistema o dispositivo adecuado (cuyo sistema de cómputo 901 es representativo) en general de un sistema de cómputo de uso general a un sistema de cómputo de uso especial modificado para facilitar el procesamiento mejorado del contenido de usuario para la identificación, anotación y ofuscación de contenido confidencial. De hecho, el *software* de codificación 905 en el sistema de almacenamiento 903 puede transformar la estructura física del sistema de almacenamiento 903. La transformación específica de la estructura física puede depender de varios factores en diferentes implementaciones de esta descripción. Los ejemplos de tales factores pueden incluir, entre otros, la tecnología utilizada para implementar los medios de almacenamiento del sistema de almacenamiento 903 y si los medios de almacenamiento informáticos se caracterizan como almacenamiento primario o secundario, así como otros factores.

Por ejemplo, si los medios de almacenamiento legibles por computadora se implementan como memoria basada en semiconductores, el *software* 905 puede transformar el estado físico de la memoria de semiconductores cuando las instrucciones de programa se codifican en ella, como al transformar el estado de transistores, condensadores, u otros elementos discretos de circuito que constituyen la memoria de semiconductores. Una transformación similar puede ocurrir con respecto a los medios magnéticos u ópticos. Son posibles otras transformaciones de medios físicos sin apartarse del alcance de la presente descripción, con los ejemplos anteriores proporcionados solo para facilitar el presente análisis.

Cada uno del entorno de aplicación DLP 906 o del entorno de DLP compartido 909 incluye uno o más elementos de *software*, como el OS 921/931 y las aplicaciones 922/932. Estos elementos pueden describir varias porciones del sistema de cómputo 901 con las cuales interactúan los usuarios, las fuentes de datos, los servicios de datos u otros elementos. Por ejemplo, el OS 921/931 puede proporcionar una plataforma de *software* en la cual se ejecuta la aplicación 922/932 y permite procesar contenido de usuario para identificación, anotación y ofuscación de contenido confidencial, entre otras funciones.

En un ejemplo, el servicio de DLP 932 incluye el repartidor de contenido 924, el anotador 925, el mapeador 926 y el ofuscador 927. El repartidor de contenido 924 reduce los elementos de contenido estructurados o jerárquicos del usuario en fragmentos lineales para su procesamiento por un servicio de clasificación. El anotador 925 resalta gráficamente datos o contenido confidenciales en una interfaz de usuario para que los usuarios puedan alertarse sobre la presencia de una cantidad umbral de datos confidenciales. El mapeador 926 puede obtener las ubicaciones específicas entre los documentos para las anotaciones de datos confidenciales, como cuando un servicio de clasificación solo proporciona intervalos/longitudes/identificaciones para localizar datos confidenciales en varios elementos estructurales o jerárquicos del documento. El ofuscador 927 presenta opciones de ofuscación para ocultar/reemplazar el contenido de usuario que se identificó como datos confidenciales. El ofuscador 927 también reemplaza el contenido confidencial que responde a las selecciones de opciones de ofuscación del usuario.

En otro ejemplo, el servicio de DLP 933 incluye el servicio de clasificación 934, el seguidor 935, el módulo de políticas/reglas 936 y el servicio de expresiones regulares 937. El servicio de clasificación 934 analiza a través de fragmentos lineales de datos o contenido para identificar datos confidenciales. El seguidor 935 mantiene conteos o cantidades de elementos de datos confidenciales encontrados por el servicio de clasificación 934 e indica los intervalos y longitudes de datos confidenciales a un mapeador para la anotación en un documento (como el mapeador 926 y el anotador 925). El módulo de políticas/reglas 936 puede recibir y mantener varias políticas y reglas para las operaciones de anotación, clasificación, detección, ofuscación u otras operaciones en el contenido de usuario. El servicio regex 937 comprende un ejemplo de técnica de clasificación que utiliza la coincidencia de expresiones regulares para identificar datos confidenciales utilizando patrones de datos o esquemas de datos, y para reemplazar el texto del contenido coincidente con contenido ofuscado.

El sistema de interfaz de comunicación 907 puede incluir conexiones de comunicación y dispositivos que permiten la comunicación con otros sistemas de cómputo (no mostrados) a través de redes de comunicación (no mostradas). Los ejemplos de conexiones y dispositivos que juntos permiten la comunicación entre sistemas pueden incluir tarjetas de interfaz de red, antenas, amplificadores de potencia, circuitos de RF, transceptores y otros circuitos de comunicación. Las conexiones y dispositivos pueden comunicarse a través de medios de comunicación para intercambiar comunicaciones con otros sistemas de cómputo o redes de sistemas, como metal, vidrio, aire o cualquier otro medio de comunicación adecuado. Los elementos físicos o lógicos del sistema de interfaz de comunicación 907 pueden recibir conjuntos de datos de fuentes de telemetría, transferir conjuntos de datos e información de control entre uno o más elementos de almacenamiento de datos distribuidos, e interactuar con un usuario para recibir

selecciones de datos y proporcionar conjuntos de datos visualizados, entre otras funciones.

El sistema de interfaz de usuario 908 es opcional y puede incluir un teclado, un ratón, un dispositivo de entrada de voz, un dispositivo de entrada táctil para recibir la entrada de un usuario. Los dispositivos de salida como una pantalla, altavoces, interfaces web, interfaces de terminal y otros tipos de dispositivos de salida también pueden incluirse en el sistema de interfaz de usuario 908. El sistema de interfaz de usuario 908 puede proporcionar la salida y recibir la entrada a través de una interfaz de red, como el sistema de interfaz de comunicación 907. En los ejemplos de red, el sistema de interfaz de usuario 908 podría empaquetar los datos de presentación o gráficos para la presentación remota mediante un sistema de presentación o sistema de cómputo conectado a través de una o más interfaces de red. Los elementos físicos o lógicos del sistema de interfaz de usuario 908 pueden recibir reglas o políticas de clasificación de parte de los usuarios o del personal de políticas, recibir la actividad de edición de datos de parte de los usuarios, presentar anotaciones de contenido confidencial a los usuarios, proporcionar opciones de ofuscación a los usuarios y presentar contenido de usuario ofuscado a los usuarios, entre otras operaciones. El sistema de interfaz de usuario 908 también puede incluir un *software* de interfaz de usuario asociado ejecutable mediante el sistema de procesamiento 902 en compatibilidad con los diversos dispositivos de entrada y salida de usuario abordados anteriormente. Por separado o en conjunto entre sí y con otros elementos de *hardware* y *software*, el *software* de interfaz de usuario y los dispositivos de interfaz de usuario pueden ser compatibles con una interfaz gráfica de usuario, una interfaz natural de usuario o cualquier otro tipo de interfaz de usuario.

La comunicación entre el sistema de cómputo 901 y otros sistemas de cómputo (no mostrados) puede ocurrir a través de una red o redes de comunicación y de acuerdo con diversos protocolos de comunicación, combinaciones de protocolos o variaciones de los mismos. Los ejemplos incluyen redes internas, redes entre redes, internet, redes de área local, redes de área amplia, redes inalámbricas, redes alámbricas, redes virtuales, redes definidas por *software*, buses de centros de datos, placas base de cómputo o cualquier otro tipo de red, combinación de red o variación de las mismas. Las redes y protocolos de comunicación previamente mencionados son bien conocidos y no necesitan abordarse en detalle en el presente documento. Sin embargo, algunos protocolos de comunicación que pueden utilizarse incluyen, entre otros, el protocolo de internet (IP, IPv4, IPv6, etcétera), protocolo de control de transmisión (TCP, por sus siglas en inglés) y protocolo de datagramas de usuario (UDP, por sus siglas en inglés), así como cualquier otro protocolo de comunicación adecuado, variación o combinación de los mismos.

Pueden apreciarse ciertos aspectos inventivos a partir de la divulgación anterior, cuyos varios ejemplos se presenta a continuación.

Ejemplo 1: Un método para proporcionar una estructura de ofuscación de datos para una aplicación de usuario, el método comprende proporcionar contenido de usuario a un servicio de clasificación configurado para procesar el contenido de usuario para clasificar las porciones del contenido de usuario como contenido confidencial correspondiente a uno o más esquemas de datos predeterminados, y recibir del servicio de clasificación las indicaciones de una o más porciones del contenido de usuario que incluye el contenido confidencial. El método incluye presentar indicaciones gráficas en una interfaz de usuario a la

aplicación de usuario que anotan una o más porciones del contenido de usuario como que incluye el contenido confidencial, y presentan opciones de ofuscación en la interfaz de usuario para ocultar el contenido confidencial dentro de al menos una porción seleccionada entre una o más porciones del contenido de usuario. En respuesta a una selección de usuario de al menos una de las opciones de ofuscación, el método incluye reemplazar el contenido de usuario asociado con contenido ofuscado que mantiene un esquema de datos del contenido de usuario asociado.

Ejemplo 2: El método del ejemplo 1, que comprende, además, presentar las opciones de ofuscación que incluyen una primera opción para ocultar el contenido confidencial dentro de la porción seleccionada y una segunda opción para ocultar el contenido confidencial dentro de la porción seleccionada y otras porciones del contenido de usuario que comprende contenido confidencial adicional que tiene un esquema de datos similar al de la porción seleccionada.

Ejemplo 3: El método de los ejemplos 1-2, que comprende, además, presentar las opciones de ofuscación que indican al menos una versión ofuscada ejemplar del contenido de usuario objetivo dentro de la porción seleccionada.

Ejemplo 4: El método de los ejemplos 1-3, donde las indicaciones gráficas que anotan una o más porciones del contenido de usuario comprenden indicadores colocados próximos a una o más porciones que pueden seleccionarse en la interfaz de usuario para presentar las opciones de ofuscación.

Ejemplo 5: El método de los ejemplos 1-4, donde el contenido ofuscado que mantiene el esquema de datos del contenido de usuario asociado comprende símbolos seleccionados en parte en el esquema de datos del contenido de usuario asociado para evitar la identificación del contenido de usuario asociado mientras se mantiene el esquema de datos del contenido de usuario asociado.

Ejemplo 6: El método de los ejemplos 1-5, que comprende, además, responder al reemplazo del contenido de usuario asociado con el contenido ofuscado, al proporcionar el contenido ofuscado al servicio de clasificación para confirmar que el contenido ofuscado no contiene contenido confidencial adicional.

Ejemplo 7: El método de los ejemplos 1-6, donde uno o más esquemas de datos predeterminados se definen mediante una o más expresiones regulares utilizadas para analizar el contenido de usuario para identificar las porciones como indicativas de uno o más patrones de contenido predeterminados o uno o más tipos de contenido predeterminados.

Ejemplo 8: El método de los ejemplos 1-7, donde uno o más de los esquemas de datos predeterminados comprende primeras porciones que deben ofuscarse y segundas porciones que permanecen no ofuscadas, las primeras porciones que deben ofuscarse corresponden a ubicaciones que tienen más de un carácter permitido, y las segundas porciones que permanecen no ofuscadas tienen solo un carácter permitido que comprende un carácter delimitador. El método comprende, además, identificar si una parte de las primeras porciones está designada para permanecer discernible por unicidad después de la ofuscación, y designar la parte para que permanezca sin ofuscar.

Ejemplo 9: Una estructura de ofuscación de datos para una aplicación de usuario, que comprende uno o más medios de almacenamiento legibles por computadora, un sistema de procesamiento conectado operativamente con uno o más medios de almacenamiento legibles por computadora, e instrucciones de

programa almacenadas en uno o más medios de almacenamiento legibles por computadora. Con base al menos en ser leída y ejecutada por el sistema de procesamiento, las instrucciones de programa dirigen al sistema de procesamiento para que al menos proporcione contenido de usuario a un servicio de clasificación configurado para procesar el contenido de usuario para clasificar porciones del contenido de usuario como que incluyen contenido confidencial correspondiente a uno o más esquemas de datos predeterminados, y reciba del servicio de clasificación indicaciones de una o más porciones del contenido de usuario como que incluyen contenido confidencial. Con base al menos en ser leídas y ejecutadas por el sistema de procesamiento, las instrucciones de programa dirigen, además, el sistema de procesamiento para que presente al menos indicaciones gráficas en una interfaz de usuario a la aplicación de usuario que anotan una o más porciones del contenido de usuario como que incluyen contenido confidencial, presente opciones de ofuscación en la interfaz de usuario para ocultar el contenido confidencial dentro de al menos una porción seleccionada entre una o más porciones del contenido de usuario, y responda a una selección de al menos una de las opciones de ofuscación, reemplace el contenido de usuario asociado con contenido ofuscado que mantiene un esquema de datos del contenido de usuario asociado.

Ejemplo 10: La estructura de ofuscación de datos del ejemplo 9, que comprende, además, instrucciones de programa, con base al menos en ser leídas y ejecutadas por el sistema de procesamiento, dirige el sistema de procesamiento para que presente al menos las opciones de ofuscación que comprenden una primera opción para ocultar el contenido confidencial dentro de la porción seleccionada y una segunda opción para ocultar el contenido confidencial dentro de la porción seleccionada y otras porciones del contenido de usuario que comprenden contenido confidencial adicional que tiene un esquema de datos similar al de la porción seleccionada.

Ejemplo 11: La estructura de ofuscación de datos de los ejemplos 9-10, que comprende, además, instrucciones de programa, con base al menos en ser leídas y ejecutadas por el sistema de procesamiento, indica al sistema de procesamiento que al menos presente las opciones de ofuscación que indican al menos un ejemplo de versión ofuscada del contenido de usuario objetivo dentro de la porción seleccionada.

Ejemplo 12: La estructura de ofuscación de datos de los ejemplos 9-11, donde las indicaciones gráficas que anotan una o más porciones del contenido de usuario comprenden indicadores colocados cerca de una o más porciones que pueden seleccionarse en la interfaz de usuario para presentar las opciones de ofuscación.

Ejemplo 13: La estructura de ofuscación de datos de los ejemplos 9-12, donde el contenido ofuscado que mantiene el esquema de datos del contenido de usuario asociado comprende símbolos seleccionados con base en parte en el esquema de datos del contenido de usuario asociado para evitar la identificación del contenido de usuario asociado mientras se mantiene el esquema de datos del contenido de usuario asociado.

Ejemplo 14: La estructura de ofuscación de datos de los ejemplos 9-13, que comprende, además, instrucciones de programa, con base al menos en ser leídas y ejecutadas por el sistema de procesamiento, dirige el sistema de procesamiento al menos, en respuesta al reemplazo del contenido de usuario asociado con el contenido ofuscado, proporcione el contenido ofuscado al servicio de clasificación para confirmar

que el contenido ofuscado no incluye otro contenido confidencial.

Ejemplo 15: La estructura de ofuscación de datos de los ejemplos 9-14, donde uno o más esquemas de datos predeterminados están definidos por una o más expresiones regulares utilizadas para analizar el contenido de usuario para identificar las porciones como indicativas de uno o más patrones de contenido predeterminados o uno o más tipos de contenido predeterminados.

Ejemplo 16: La estructura de ofuscación de datos de los ejemplos 9-15, donde uno o más esquemas de datos predeterminados comprenden primeras porciones que se ofuscarán y segundas porciones para permanecer no ofuscadas, las primeras porciones que se ofuscarán corresponden a ubicaciones que tienen más de un carácter permitido, y las segundas porciones que permanecen no ofuscadas tienen solo un carácter permitido que comprende un carácter delimitador. La estructura de ofuscación de datos que comprende, además, instrucciones de programa, con base al menos en ser leídas y ejecutadas por el sistema de procesamiento, dirige el sistema de procesamiento para que al menos identifique si una parte de las primeras porciones están designadas para permanecer discernibles por unidad después de la ofuscación, y designe la parte que permanecerá sin ofuscar.

Ejemplo 17: Un método para operar una aplicación de usuario, el método comprende proporcionar contenido de usuario de un archivo de datos de usuario a un servicio de clasificación configurado para procesar el contenido de usuario para clasificar una o más porciones del contenido de usuario como que comprenden contenido confidencial correspondiente a uno o más esquemas de datos, y presentar indicadores en una interfaz de usuario que señalan que una o más porciones del contenido de usuario incluyen contenido confidencial, donde los indicadores se colocan próximos a una o más porciones y pueden seleccionarse en la interfaz de usuario para presentar opciones de ofuscación. En respuesta a la selección de uno de los primeros indicadores, el método incluye presentar las primeras opciones de ofuscación en la interfaz de usuario para reemplazar el contenido confidencial asociado dentro de una primera porción del contenido de usuario marcado por el primero de los indicadores. En respuesta a una selección de usuario de al menos una de las primeras opciones de ofuscación, el método incluye reemplazar el contenido confidencial asociado con contenido ofuscado que mantiene un esquema de datos del contenido confidencial asociado.

Ejemplo 18: El método del ejemplo 17, que comprende, además, presentar las primeras opciones de ofuscación que comprenden una primera opción para reemplazar el contenido confidencial asociado con el contenido ofuscado y una segunda opción para reemplazar el contenido confidencial asociado y el contenido confidencial adicional del archivo de datos de usuario que tiene un esquema de datos similar al contenido confidencial asociado.

Ejemplo 19: El método de los ejemplos 17-18, donde el contenido ofuscado que mantiene el esquema de datos del contenido confidencial asociado comprende uno o más símbolos seleccionados para evitar la identificación del contenido confidencial asociado mientras se mantiene el esquema de datos del contenido de usuario asociado, donde uno o más símbolos se seleccionan en parte en el esquema de datos del contenido confidencial asociado.

Ejemplo 20: El método de los ejemplos 17-19, donde uno o más esquemas de datos comprenden

primeras porciones que se ofuscarán y segundas porciones que no se ofuscarán, las primeras porciones que se ofuscarán corresponden a ubicaciones que tienen más de un carácter permitido, y las segundas porciones que permanecerán no ofuscadas tienen solo un carácter permitido que comprende un carácter delimitador. El método comprende, además, identificar si una parte de las primeras porciones está designada para permanecer discernible por unicidad después de la ofuscación, y designar la parte para que permanece sin ofuscar.

Los diagramas de bloques funcionales, casos y secuencias operacionales, y diagramas de flujo proporcionados en las figuras son representativos de sistemas, entornos y metodologías ejemplares para realizar aspectos novedosos de la divulgación. Aunque, para simplificar la explicación, los métodos incluidos en el presente documento pueden tener la forma de un diagrama funcional, caso o secuencia operativa, o diagrama de flujo, y pueden describirse como una serie de acciones, debe entenderse y apreciarse que los métodos no están limitados por el orden de las acciones, puesto que algunas acciones pueden, de acuerdo con las mismas, ocurrir en un orden diferente y/o concurrentemente con otras acciones de aquellas mostradas y descritas en el presente documento. Por ejemplo, los técnicos en la materia entenderán y apreciarán que un método podría representarse alternativamente como una serie de estados o eventos interrelacionados, como en un diagrama de estados. Además, no todas las acciones ilustradas en una metodología pueden ser necesarios para una implementación novedosa.

Las descripciones y figuras incluidas en el presente documento representan implementaciones específicas para enseñar a los técnicos en la materia cómo llevar a cabo y utilizar la mejor opción. A fin de enseñar los principios inventivos, algunos aspectos convencionales se han simplificado u omitido. Los técnicos en la materia apreciarán las variaciones de estas implementaciones que se encuentran dentro del alcance de la divulgación. Los técnicos en la materia también apreciarán que las funciones descritas anteriormente pueden combinarse de varias maneras para formar múltiples implementaciones. Como resultado, la invención no se limita a las implementaciones específicas descritas anteriormente, sino solo a las reivindicaciones y sus equivalentes.