(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 755 924**

(51) Int. CI.:

G06F 16/178 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **17.03.2015 E 15159520 (4)**

(97) Fecha y número de publicación de la concesión europea: **21.08.2019 EP 2921975**

(54) Título: **Determinación y extracción de datos modificados procedentes de una fuente de datos**

(30) Prioridad:

18.03.2014 US 201461955054 P
16.04.2014 US 201414254757
16.04.2014 US 201414254773

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
24.04.2020

(73) Titular/es:

PALANTIR TECHNOLOGIES INC. (100.0%)
100 Hamilton Avenue
Palo Alto, CA 94301 , US

(72) Inventor/es:

FISHER, WILLIAM y
MAAG, PETER

(74) Agente/Representante:

TORNER LASALLE, Elisabet

DESCRIPCIÓN

Determinación y extracción de datos modificados procedentes de una fuente de datos

Solicitudes de prioridad

La presente solicitud reivindica el beneficio de la solicitud U.S. nº 14/254.757, presentada el 16 de abril de 2014, de la solicitud U.S. nº 14/254.773, presentada el 16 de abril de 2014 y de la solicitud provisional U.S. nº 61/955054, presentada el 18 de marzo de 2014.

Campo técnico

La presente divulgación versa acerca de sistemas y de técnicas para la integración y el análisis de datos. Más específicamente, la presente divulgación versa acerca de la identificación de modificaciones en los datos de una fuente de datos.

Antecedentes

Las organizaciones y/o empresas producen cantidades cada vez más grandes de datos. Tales datos pueden ser almacenados en distintas fuentes de datos. Las fuentes de datos pueden ser actualizadas, por ejemplo, de forma periódica.

El documento WO 2013/030595 A1 da a conocer un sistema para identificar unidades de datos con el fin de sincronizar las unidades de datos entre una pluralidad de dispositivos en función de una asignación de un testigo de actualización que es revocado por un sucesivo testigo de actualización.

Sumario

La invención está definida por las reivindicaciones independientes, teniendo en cuenta debidamente cualquier elemento que sea equivalente a un elemento especificado en las reivindicaciones. Las reivindicaciones dependientes versan acerca de características opcionales de algunas realizaciones de la invención.

Cada uno de los sistemas, procedimientos y dispositivos descritos en la presente memoria tiene varios aspectos, ninguno de los cuales es responsable único de sus atributos deseables. Sin limitar el alcance de la presente divulgación, se expondrán ahora brevemente varias características no limitantes.

En una realización, un sistema de ordenador configurado para obtener datos modificados de una fuente de datos comprende: uno o más procesadores de ordenador de soporte físico configurados para ejecutar código para provocar que el sistema: obtenga información que indica una pluralidad de agrupamientos de datos almacenados en uno o más ficheros o bases de datos en una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determine un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han cambiado comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un primer número histórico de unidades de datos incluidas en una versión local correspondiente del primer agrupamiento, en el que se crea la versión local correspondiente del primer agrupamiento en función de las unidades de datos incluidas en el primer agrupamiento en un primer momento antes de dicha obtención de la información que indica la pluralidad de agrupamientos de los datos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de la versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

En otra realización, un procedimiento para obtener datos modificados de una fuente de datos comprende: obtener, mediante uno o más procesadores de ordenador de soporte físico, información que indica una pluralidad de agrupamientos de datos almacenados en uno o más ficheros o bases de datos en una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determinar, mediante los uno o más procesadores de ordenador de soporte físico, un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han cambiado comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un primer número histórico de unidades de datos incluidas en una versión local correspondiente del primer agrupamiento, en el que se crea la versión local correspondiente del primer agrupamiento en función de las unidades de datos incluidas en el primer agrupamiento en un primer momento antes de dicha obtención de la información que indica la pluralidad de agrupamientos de los datos; acceder, mediante los uno o más procesadores de ordenador de soporte físico, unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; comparar, mediante los uno o más procesadores de ordenador de soporte físico, las unidades de datos incluidas en el primer agrupamiento con unidades de datos de la versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas; extraer, mediante los uno o más procesadores de ordenador de soporte físico, las unidades modificadas de datos del primer agrupamiento; y remitir, mediante los

uno o más procesadores de ordenador de soporte físico, las unidades modificadas extraídas de datos a un sistema de destino.

En otra realización más, un medio transitorio o no transitorio legible por un ordenador comprende instrucciones para obtener datos modificados procedentes de una fuente de datos que provocan que un procesador de ordenador: obtenga información que indica una pluralidad de agrupamientos de datos almacenados en uno o más ficheros o bases de datos en una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determinar un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un primer número histórico de unidades de datos incluidas en una versión local correspondiente del primer agrupamiento, en el que se crea la versión local correspondiente del primer agrupamiento en función de unidades de datos incluidas en el primer agrupamiento en un primer momento antes de dicha obtención de la información que indica la pluralidad de agrupamientos de los datos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de la versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

En algunas realizaciones, un sistema de ordenador configurado para obtener datos modificados procedentes de una fuente de datos comprende: uno o más procesadores de ordenador de soporte físico configurados para ejecutar código para provocar que el sistema: obtenga información que indica una pluralidad de agrupamientos de datos de una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determine un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un número histórico de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de una versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas, en el que la versión local correspondiente del primer agrupamiento de unidades de datos es una versión comprimida del primer agrupamiento de unidades de datos; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

En ciertas realizaciones, un procedimiento para obtener datos modificados procedentes de una fuente de datos comprende: obtener, mediante uno o más procesadores de ordenador de soporte físico, información que indica una pluralidad de agrupamientos de datos de una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determinar, mediante los uno o más procesadores de ordenador de soporte físico, un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un número histórico de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; acceder, mediante los uno o más procesadores de ordenador de soporte físico, a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; comparar, mediante los uno o más procesadores de ordenador de soporte físico, las unidades de datos incluidas en el primer agrupamiento con unidades de datos de una versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas, en el que la versión local correspondiente del primer agrupamiento de unidades de datos es una versión comprimida del primer agrupamiento de unidades de datos; extraer, mediante los uno o más procesadores de ordenador de soporte físico, las unidades modificadas de datos del primer agrupamiento; y remitir, mediante los uno o más procesadores de ordenador de soporte físico, las unidades modificadas extraídas de datos a un sistema de destino.

En otras realizaciones, un medio transitorio o no transitorio legible por un ordenador comprende instrucciones para obtener datos modificados procedentes de una fuente de datos que provoca que un procesador de ordenador: obtenga información que indica una pluralidad de agrupamientos de datos de una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determine un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un número histórico de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de una versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas, en el que la versión local correspondiente del primer agrupamiento de unidades de datos es una versión comprimida del primer agrupamiento de unidades de datos; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

Breve descripción de los dibujos

La Figura 1 es un diagrama de bloques que ilustra una realización de un sistema de determinación de modificaciones configurado para determinar y obtener modificaciones en los datos de una pluralidad de fuentes de datos.

La Figura 2A es un diagrama de flujo de datos ilustrativo de la interacción entre los diversos componentes de un sistema de determinación de modificaciones configurados para determinar y obtener las modificaciones en los datos de una pluralidad de fuentes de datos, según una realización.

La Figura 2B es un diagrama de flujo de datos ilustrativo de la interacción entre los diversos componentes de un sistema de determinación de modificaciones configurados para determinar y obtener las modificaciones en los datos de una pluralidad de fuentes de datos, según otra realización.

La Figura 3 es un ejemplo de información obtenida de una fuente de datos y/o de información procesada por el sistema de determinación de modificaciones.

La Figura 4A es un diagrama de flujo que ilustra una realización de un procedimiento para determinar y obtener modificaciones en los datos de una pluralidad de fuentes de datos.

La Figura 4B es un diagrama de flujo que ilustra otra realización de un procedimiento para determinar y obtener modificaciones en los datos de una pluralidad de fuentes de datos.

La Figura 5 es un diagrama de bloques que ilustra un sistema de ordenador con el que pueden implementarse ciertos procedimientos expuestos en la presente memoria.

Descripción detallada

Visión general

Las empresas pueden necesitar obtener datos de una o más fuentes de datos. A menudo, los datos para un marco temporal particular son descargados de una fuente de datos. Por ejemplo, una fuente de datos puede contener ficheros de registros, y se pueden descargar ficheros de registros de los últimos dos días. Sin embargo, algunos de los datos pueden haber sido obtenidos ya en un momento anterior, y es posible que el sistema que está solicitando los datos no pueda distinguir entre datos que ya tiene y datos nuevos o modificados que aún no ha obtenido. Por ejemplo, el sistema peticionario puede simplemente almacenar los datos que descarga cada vez sin considerar si se duplican algunos datos. Algunos datos pueden ser descargados de nuevo aunque ya existan en el sistema. En consecuencia, existe la necesidad de identificar ni de extraer datos modificados de una fuente de datos de una forma eficaz.

Según se divulga en la presente memoria, un sistema de determinación de modificaciones puede estar configurado para identificar y obtener las modificaciones en los datos procedentes de una o más fuentes de datos. Por ejemplo, el sistema puede determinar que existen modificaciones a los datos de una fuente de datos (por ejemplo, los datos para un marco temporal particular, tal como un día) en función de cierta información de resumen para un conjunto actual de datos de la fuente de datos (por ejemplo, líneas de datos asociadas con un día particular en un conjunto de datos recibido anteriormente) en comparación con un conjunto actual de datos de la fuente de datos (por ejemplo, líneas de datos asociadas con el día particular en un conjunto actual de datos). Una vez se identifican los datos con modificaciones (por ejemplo, días con distintas cantidades de líneas de datos), se pueden obtener los datos modificados y ser comparados con una versión local de los datos (o alguna representación de los datos) para identificar las unidades particulares de datos (por ejemplo, líneas particulares de datos) que han sido modificadas, de forma que la fuente de datos solo necesite proporcionar esas unidades particulares de datos.

Una fuente de datos puede ser una o más bases de datos y/o uno o más ficheros. Las modificaciones reales pueden ser remitidas a un sistema de destino para su almacenamiento. El sistema de determinación de modificaciones puede actuar como un intermediario entre fuentes de datos y uno o más sistemas de destino para identificar datos modificados y remitir únicamente los datos modificados a los sistemas de destino.

Puede llevar mucho tiempo una descarga de una fuente de datos (por ejemplo, debido a la velocidad reducida, la cantidad de datos, etc.), y la nueva descarga de datos que ya existen en el sistema de destino puede dar lugar a un gasto de tiempo y de recursos innecesarios. Además, el almacenamiento de datos duplicados puede ocupar un espacio innecesario de almacenamiento en el sistema de destino. Al identificar y remitir únicamente los datos modificados, el sistema de determinación de modificaciones puede proporcionar una forma de obtener datos de una fuente de datos de una forma inteligente y puede ahorrar tiempo y/o recursos para el sistema de destino. Esto puede ser muy útil especialmente cuando una fuente de datos contiene grandes cantidades de datos, y únicamente se ha modificado una porción reducida de los datos. El sistema de determinación de modificaciones también puede

identificar las modificaciones rápidamente, por ejemplo, llevando a cabo una operación de agrupamiento de los datos explicada en detalle a continuación.

Sistema de determinación de modificaciones

La Figura 1 es un diagrama de bloques que ilustra una realización de un sistema 100 de determinación de modificaciones configurado para determinar y obtener las modificaciones en los datos de una pluralidad de fuentes 110 de datos. Una fuente 110 de datos puede incluir una o más bases 110a de datos, uno o más ficheros 110b (por ejemplo, un fichero o un sistema de ficheros planos), cualquier otro tipo de estructura de datos, o una combinación de múltiples estructuras de datos. Los datos en la base 110a de datos pueden organizarse en una o más tablas, que incluyen filas y columnas. Los datos en el fichero 110b pueden organizarse como líneas con diversos campos. Por ejemplo, un fichero 110b puede tener un formato CSV. Se puede hacer referencia a los datos en los ficheros, en las bases de datos o en cualquier otra estructura de datos, en términos de “líneas”, siendo una línea un subconjunto del fichero. Por ejemplo, las líneas de un fichero pueden ser grupos de datos entre marcas de salto de línea o divisiones de texto del fichero en grupos de tamaño predeterminado (por ejemplo, cada línea incluye 255 caracteres) y las líneas de una base de datos pueden ser una fila o algún otro subconjunto de información en la base de datos. Los datos en una base 110a de datos y en un fichero 110b pueden ser gestionados o procesados de una forma similar por el sistema 100 de determinación de modificaciones (“CDS”). En ciertas realizaciones, el CDS 100 obtiene las modificaciones de una única fuente 110 de datos, en vez de de múltiples fuentes 110 de datos.

El CDS 100 puede incluir uno o más componentes (no mostrados) que llevan a cabo funciones relacionadas con la determinación y la obtención de modificaciones en los datos de las fuentes 110 de datos. El CDS 100 también puede incluir un almacenamiento local 150, que puede almacenar cualquier versión local de los datos en una fuente 110 de datos, tal como una versión resumida y/o comprimida de los datos. La versión local de los datos puede ser utilizada para identificar las modificaciones reales en los datos de la fuente 210 de datos y/o las porciones de los datos (por ejemplo, líneas de los datos) que incluyen modificaciones. La versión local puede incluir algunos de los datos, o todos ellos, de una fuente de datos, dependiendo de la realización.

Uno o más sistemas 270 de destino pueden solicitar datos modificados del CDS 100. Los sistemas 270 de destino pueden enviar una solicitud de forma periódica (por ejemplo, programada), bajo petición, etc. El CDS 100 también puede comprobar periódicamente una o más fuentes 210 de datos y remitir cualquier modificación sin recibir una solicitud procedente de un sistema 270 de destino. Por ejemplo, se puede programar el CDS 100 para que compruebe las fuentes 210 de datos cada 2 horas. Tal programación puede definirse como una o más directrices. En una realización, un sistema 270 de destino incluye el CDS 100, de forma que se pueda llevar a cabo la funcionalidad descrita en la presente memoria con referencia al CDS 100 por medio del propio sistema 270 de destino.

La Figura 2A es un diagrama de flujo de datos ilustrativo de la interacción entre los diversos componentes de un sistema 200a de determinación de modificaciones configurado para determinar y obtener las modificaciones en los datos de una pluralidad de fuentes de datos, según una realización. El CDS 200a y los componentes correspondientes de la Figura 2A pueden ser similares o iguales al CDS 100 y a los componentes denominados de forma similar de la Figura 1.

En la acción 1 de flujo de datos, el CDS 200a lleva a cabo una consulta en los datos en una fuente 210 de datos agrupar por un atributo particular, tal como una columna de información en una tabla. Con fines de la exposición de la presente memoria, se exponen muchos ejemplos con referencia a un agrupamiento en función de una o más “columnas”, estando asociada cada columna con un atributo particular. En otras realizaciones, los atributos pueden asociarse con distintas características de visualización de una estructura de datos (por ejemplo, además de columnas). Por ejemplo, si la fuente 210 de datos es una base 210a de datos, el CDS 200a puede llevar a cabo una consulta SQL y agrupar por una columna particular en una tabla, por ejemplo, utilizando la cláusula AGRUPAR POR, que se utiliza en SQL para agrupar filas que tienen valores comunes en un conjunto más pequeño de filas. Los conjuntos más pequeños de filas pueden ser denominados particiones o grupos. Cada partición incluye filas que tienen el mismo valor para la columna designada. A menudo se utiliza AGRUPAR POR junto con funciones de agregación SQL o para eliminar filas duplicadas de un conjunto de resultados.

La columna que se designa como la columna para AGRUPAR POR debería poder proporcionar alguna indicación de qué filas son nuevas o modificadas con respecto a la anterior vez que el CDS 200a obtuvo datos de la fuente 210 de datos. En un ejemplo, una tabla en la base 210a de datos incluye una columna actualizada por última vez, que incluye un sello de tiempo de cuando se actualizaron por última vez los datos en la fila, y los datos en la base 210a de datos pueden ser agrupados por la columna actualizada por última vez. Dado que el sello de tiempo puede incluir la hora, el minuto, el segundo, etc. además de la fecha, solo la fecha del sello de tiempo podría ser utilizada para AGRUPAR POR. En tal caso, una partición estaría basada en un día, y cada partición contendría las filas de cada día. Se puede aplicar una función de agregación, tal como RECUENTO a los resultados de AGRUPAR POR para obtener el número de filas para cada partición. Los resultados de la consulta de la fuente 210 de datos pueden incluir una o más particiones del AGRUPAR POR y el número de filas incluidas en cada partición. El CDS 200a puede

almacenar los resultados o hacer un seguimiento de los resultados localmente, de forma que se puedan comparar los resultados con la siguiente vez que el CDS 200a solicita este tipo de información de la fuente 210 de datos.

Los datos de una fuente 210b de datos de fichero también pueden ser consultados de una forma similar. El CDS 200a puede utilizar distintos adaptadores para acceder a los datos que residen en una o más bases 210a de datos y a los datos que residen en uno o más ficheros 210b, pero una vez se obtienen los datos, pueden ser gestionados de la misma forma por el CDS 200a, con independencia de si la fuente de datos tiene la forma de una base 210a de datos o de un fichero 210b. Los detalles expuestos con respecto a una fuente 210a de datos de base de datos pueden generalizarse a otros tipos de fuentes 210 de datos, incluyendo una fuente 210b de datos de fichero. Por ejemplo, una partición puede hacer referencia a un agrupamiento utilizado en un texto (u otro tipo de campo) resultante de una operación que es similar al SQL AGRUPAR POR. También se puede denominar a una partición un "agrupamiento". Un agrupamiento puede incluir datos de una base 210a de datos o de un fichero 210b. Se puede denominar "unidad de datos" a una unidad de datos incluida en un agrupamiento. Una unidad de datos puede ser una fila en el caso de una base 210a de datos o una línea en el caso de un fichero 210b. La columna o el campo agrupar por puede ser cualquier tipo que pueda proporcionar particiones de tamaño apropiado para la comparación (por ejemplo, proporcionar una distribución no homogénea o no uniforme). A continuación se explican algunos detalles relativos al grupo por columna/campo.

En la acción 2 de flujo de datos, el CDS 200a determina qué partición o particiones han sido modificadas. El número de filas para las particiones obtenidas en la acción 1 de flujo de datos puede compararse con el número de filas para particiones correspondientes obtenidas en un momento anterior. El número actual de filas en particiones puede ser denominado "datos actuales de agrupamiento", y el número de filas para diversas particiones de un momento anterior puede ser denominado "datos históricos de agrupamiento". El CDS 200a puede comparar los datos actuales de agrupamiento con los datos históricos de agrupamiento para determinar si se modificó el número de filas para una partición particular. Por ejemplo, si el número de filas para el Día 1 es 1.000 en los datos históricos de agrupamiento, pero el número de filas para el Día 1 es 1.050 en los datos actuales de agrupamiento, la partición para el Día 1 es un candidato para verificar si se modificaron los datos reales. Es probable que se añadieran 50 nuevas filas para el Día 1, y el CDS 200a puede determinar qué filas de las 1.050 son nuevas y extraerlas para remitirlas a un sistema 270 de destino. De esta forma, el CDS 200a puede identificar una o más particiones que han sido modificadas.

En la acción 3 de flujo de datos, el CDS 200a obtiene datos para cualquier partición modificada identificada. En particular, una vez que el CDS 200a identifica una partición modificada, el CDS 200a obtiene los datos para la partición particular de la fuente 210 de datos. En el anterior ejemplo, el CDS 200a solicita las 1.050 filas para el Día 1. El CDS 200a puede almacenar las 1.050 filas localmente, por ejemplo, en el almacenamiento local 250a. Entonces, se pueden utilizar las 1.050 filas para una comparación la siguiente vez que se modifican los datos para esta partición.

En la acción 4 del flujo de datos, el CDS 200a compara los datos obtenidos con la versión local de la partición modificada. El CDS 200a puede comparar los datos descargados con una versión local correspondiente de los datos. Por ejemplo, el CDS 200a puede haber almacenado las 1.000 filas para el Día 1 a partir de un momento anterior en el almacenamiento local 250a. Los datos para la partición de un momento anterior pueden ser denominados "datos históricos de partición". De forma similar, los datos actuales para la partición pueden ser denominados "datos actuales de partición". La versión local de los datos puede incluir datos para una partición o un número de particiones. Al comparar las 1.050 filas actuales para el Día 1 en los datos actuales de partición con las 1.000 filas anteriores para el Día 1 en los datos históricos de partición, el CDS 200a puede identificar qué 50 filas han sido modificadas. Estas 50 filas pueden ser marcadas o colocadas en un fichero, de forma que puedan ser enviadas a cualquier sistema relevante 270 de destino.

En la acción 5 de flujo de datos, el CDS 200a remite los datos modificados al sistema 270 de destino. Según se ha explicado anteriormente, los datos modificados pueden ser extraídos de los datos actuales de partición, guardados en un fichero, y enviados al sistema 270 de destino. Entonces, el sistema 270 de destino puede almacenar los datos recibidos en su o sus dispositivos de almacenamiento sin tener que comprobar si existe cualquier duplicado en los datos recibidos. Debido a que el CDS 210a puede enviar los datos modificados exactos, el sistema 270 de destino puede simplemente almacenar lo que recibe del CDS 210a y no necesita implementar mucha funcionalidad en su extremo.

En la Figura 2A, para compara los datos actuales de partición con los datos históricos de partición, el CDS 200a puede mantener localmente todos los datos, o un subconjunto de los mismos, de una fuente 210 de datos. Sin embargo, en algunos casos, el almacenamiento local 250a puede tener un espacio limitado de almacenamiento, y el CDS 200a puede no poder almacenar todos los datos utilizados en la comparación. En consecuencia, en tales casos, el CDS 200a puede mantener una versión comprimida de los datos procedentes de una fuente 210 de datos para su comparación. En algunas realizaciones, la versión comprimida de los datos puede ser uno o más filtros de Bloom. Tales realizaciones se describen con más detalle en conexión con la Figura 2B.

La Figura 2B es un diagrama de flujo de datos ilustrativo de la interacción entre los diversos componentes de un sistema 200b de determinación de modificaciones configurado para determinar y obtener las modificaciones en datos de una pluralidad de fuentes de datos, según otra realización. El CDS 200b y los componentes correspondientes de la Figura 2B pueden ser similares o iguales que el CDS 100, 200a y los componentes denominados de forma similar de las Figuras 1 y 2A.

Las acciones 1-5 de flujo de datos pueden ser similares a las acciones 1-5 de flujo de datos en la Figura 2A. Anteriormente se han explicado con más detalle ciertos detalles relativos al CDS 200b en conexión con la Figura 2A. En general, no obstante, en la acción 1 de flujo de datos, el CDS 200b lleva a cabo una consulta de los datos en una fuente 210 de datos para agrupar por una columna particular. En la acción 2 de flujo de datos, el CDS 200b determina qué partición ha sido modificada. En la acción 3 de flujo de datos, el CDS 200b obtiene datos para la partición modificada.

Según se ha explicado anteriormente, se pueden utilizar uno o más filtros de Bloom en comparación con los datos actuales de partición y con los datos históricos de partición. Un filtro de Bloom puede hacer referencia a una estructura probabilística de datos que ahorra espacio que se utiliza para comprobar si un elemento es un miembro de un conjunto (por ejemplo, las unidades de datos son parte de una partición). Son posibles coincidencias de falsos positivos, pero no lo son los falsos negativos. Por ejemplo, una consulta puede devolver bien “posiblemente en el conjunto” o “definitivamente no en el conjunto”. Se pueden añadir elementos al conjunto, pero, en general, no pueden ser eliminados. Según se añaden más elementos al conjunto, la probabilidad de falsos positivos se vuelve mayor.

En una realización, un filtro de Bloom utilizado por el CDS 200b es una matriz de bits de m bits y tiene k distintas funciones de comprobación aleatoria que se utilizan para añadir un elemento al filtro de Bloom. Para añadir un elemento al filtro de Bloom, se suministra un elemento a cada una de las k funciones de comprobación aleatoria para obtener k posiciones de matriz. Los bits en estas k posiciones están puestos a 1. Para consultar un elemento para determinar si se encuentra en el conjunto, se suministra el elemento a cada una de las k funciones de comprobación aleatoria para obtener las k posiciones de matriz. Si cualquier de los bits en estas posiciones es 0, el elemento no está definitivamente en el conjunto. Si todos los bits en estas posiciones son 1, el elemento se encuentra bien en el conjunto, o los bits fueron puestos a 1 por casualidad cuando se añadieron otros elementos. Si los bits fueron puestos a 1 por casualidad, esto puede dar lugar a un falso positivo. No se requiere el filtro de Bloom para almacenar los propios elementos.

Aunque existe un riesgo de falsos positivos, los filtros de Bloom pueden proporcionar una gran ventaja de espacio con respecto a otras estructuras de datos para representar conjuntos, tales como árboles autorregulados de búsqueda binaria, tablas de direcciones calculadas, matrices sencillas, listas enlazadas, etc. Otras estructuras de datos pueden requerir almacenar al menos las propias unidades de datos, lo que puede requerir una longitud cualquiera desde un número reducido de bits (por ejemplo, para números enteros pequeños) hasta un número arbitrario de bits (por ejemplo, para cadenas). Por otra parte, un filtro de Bloom con un error de 1% y un valor óptimo de k puede requerir únicamente aproximadamente 9,6 bits por elemento (por ejemplo, unidad de datos), con independencia del tamaño de los elementos. La ventaja de espacio puede ser parcialmente debida a la compacidad del filtro de Bloom, heredada de las matrices, y parcialmente debida a la naturaleza probabilística del filtro de Bloom. Se puede reducir la tasa de falsos positivos de 1% en un factor de diez añadiendo únicamente aproximadamente 4,8 bits por elemento.

En la acción 4 de flujo de datos, el CDS 200b compara los datos actuales de partición con datos históricos de partición utilizando uno o más filtros 255 de Bloom. Se pueden almacenar uno o más filtros 255 de Bloom en un almacenamiento local 250b. Se puede crear un filtro de Bloom para la versión local de los datos. Por ejemplo, en el momento de una descarga anterior, el CDS 200b pudo haber añadido las unidades de datos procedentes de la fuente 210 de datos a un filtro de Bloom. Aunque el filtro de Bloom no almacena las unidades reales de datos (por ejemplo, las unidades reales de datos pueden ser borradas después de que se generan los filtros de Bloom correspondientes), puede determinar con una probabilidad elevada si se incluyó una unidad de datos en la anterior versión de los datos o no. El filtro de Bloom puede ocupar mucho menos espacio que el almacenamiento de los datos históricos de partición o actuales y puede servir de versión comprimida de los datos. Para cada unidad de datos incluida en los datos actuales de partición, el CDS 200b puede consultar el filtro de Bloom que incluye los datos históricos de partición correspondientes para comprobar si se incluyó la unidad de datos en los datos históricos de partición o si es nueva. En una realización, si se definen n número de funciones de comprobación aleatoria para el filtro de Bloom, el filtro de Bloom aplica las n funciones de comprobación aleatoria a la unidad de datos para devolver n número de posiciones de matriz. Si cualquiera de las posiciones de matriz es 0, no se incluyó la unidad de datos en los datos históricos de partición. Si todas las posiciones de matriz son 1, probablemente se incluyó la unidad de datos en los datos históricos de partición, aunque existe una pequeña probabilidad de falsos positivos.

Los datos de partición obtenidos de una fuente 210 de datos pueden incluir un número de unidades individuales de datos (por ejemplo, filas, líneas, etc.), y el número real de unidades de datos incluidas en los datos de una partición puede variar; algunas particiones pueden incluir un número pequeño de unidades de datos, y otras particiones

pueden incluir un gran número de unidades de datos. En una realización, el tamaño de un filtro de Bloom es predeterminado, y puede no ser óptimo utilizar el mismo filtro de Bloom para una cantidad reducida de datos y para una gran cantidad de datos. La probabilidad de que el filtro de Bloom devuelva falsos positivos aumenta con el número de elementos añadidos al filtro de Bloom. Por lo tanto, si se añaden demasiados elementos, el filtro de Bloom puede quedar saturado y la precisión del filtro de Bloom puede deteriorarse, por ejemplo, hasta un punto de devolver casi un 100% de falsos positivos. En consecuencia, se pueden utilizar los filtros de Bloom de distintos tamaños para acomodar datos de tamaño variable. Por ejemplo, un filtro de Bloom tiene un tamaño predeterminado de m bits cuando se crea y es posible que no pueda acomodar datos que incluyan más de un número específico de elementos (por ejemplo, x número de elementos) sin un deterioro de la precisión. Para datos que incluyen más de x elementos, se puede utilizar un filtro de Bloom que tiene un tamaño mayor que m bits. Debido a que puede variar el tamaño de los datos de distintas fuentes de datos, el CDS 200b puede utilizar una serie de filtros de Bloom de tamaño creciente para acomodar distintos tamaños de datos. Por ejemplo, el CDS 200b puede tener un número de filtros de Bloom de tamaños variables disponibles para ser utilizados, o puede crear uno según sea necesario. En un ejemplo, el CDS 200b puede comenzar con un filtro de Bloom que tiene un tamaño de m bits, y si este filtro de Bloom es demasiado pequeño para los datos, el CDS 200b puede seleccionar o crear un filtro de Bloom que tenga un tamaño de m + y bits, etcétera, hasta que el CDS 200b encuentre un filtro de Bloom que tenga el tamaño correcto para los datos. Los datos de diversas fuentes 210 de datos pueden compartir el mismo conjunto de filtros de Bloom. O en ciertas realizaciones, el CDS 200b puede mantener los filtros de Bloom para distintas fuentes 210 de datos separados entre sí.

En una realización, el CDS 200b puede almacenar los filtros 255 de Bloom en el almacenamiento que proporciona una accesibilidad elevada. Por ejemplo, el almacenamiento local 250b puede aumentar el almacenamiento que es un tipo que es más accesible que el almacenamiento utilizado por un sistema 270 de destino. Por ejemplo, el almacenamiento local 250b puede utilizar un almacenamiento conectado en red (NAS), dado que es muy accesible a dispositivos conectados. Un tipo más accesible de almacenamiento puede ser más costoso que un tipo de almacenamiento menos accesible, y dado que los filtros de Bloom pueden ahorrar espacio, el CDS 200b puede reducir los costes asociados con el almacenamiento local 250b.

En la acción 5 de flujo de datos, el CDS 200b extrae y remite los datos modificados al sistema 270 de destino. Esta etapa puede ser similar a la acción 5 de flujo de datos de la Figura 2A. El CDS 200b puede remitir los datos modificados a uno o más sistemas 270 de destino.

La Figura 3 es un ejemplo de información obtenido de una fuente de datos y/o información procesada por el sistema de determinación de modificaciones. Se explicará un ejemplo ilustrativo específico con respecto a la Figura 3. Se explicarán diversos aspectos con referencia al CDS 200a en la Figura 2A, pero el ejemplo también se puede aplicar al CDS 100, 200b de las Figuras 1 y 2B. El ejemplo hará referencia a datos de una fuente 210 de datos en el instante T0 y a datos de la fuente 210 de datos en el instante T1, siendo T0 anterior a T1.

En el instante T1, el CDS 200a lleva a cabo una consulta en los datos de la fuente 210 de datos para agrupar los datos por la columna o el campo actualizado o modificado por última vez. Los datos pueden ser agrupados en uno o más agrupamientos en función de la fecha. La fuente 210 de datos puede devolver un resultado que incluye agrupamientos 310 organizados por fecha. El resultado puede denominarse “datos actuales de agrupamiento”. Los datos actuales 310 de agrupamiento pueden enumerar la fecha para un agrupamiento y el número de unidades de datos incluidos en ese agrupamiento. Los datos actuales 310 de agrupamiento muestran que el Agrupamiento 1 es para 10/2/14, y el número de unidades de datos en el Agrupamiento 1 es de 400; el Agrupamiento 2 es para 11/2/14, y el número de unidades de datos en el Agrupamiento 2 es de 310; y el Agrupamiento 3 es para 12/2/14, y el número de unidades de datos en el Agrupamiento 3 es de 175.

El CDS 200a compara los datos actuales 310 de agrupamiento con los datos históricos 315 de agrupamiento. Los datos históricos 315 de agrupamiento pueden incluir los datos de agrupamiento obtenidos de la fuente 210 de datos en diversos momentos del pasado. Los datos históricos 315 de agrupamiento pueden incluir datos de agrupamiento para uno o más días. En la Figura 3, los datos históricos 315 de agrupamiento muestran los datos de agrupamiento en T0. Los datos históricos 315 de agrupamiento muestran que el Agrupamiento 1 es para 10/2/14, y el número de unidades de datos en el Agrupamiento 1 es de 380; el Agrupamiento 2 es para 11/2/14, y el número de unidades de datos en el Agrupamiento 2 es de 310; y el Agrupamiento 3 es para 12/2/14, y el número de unidades de datos en el Agrupamiento 3 es de 165.

Al comparar el número de unidades de datos en los mismos agrupamientos en distintos instantes, el CDS 200 puede identificar que ciertos agrupamientos han sido modificados o son candidatos potenciales que tienen unidades modificadas de datos. El número de unidades de datos para el Agrupamiento 1 en T1 es de 400, y el número de unidades de datos para el Agrupamiento 1 en T0 es de 380. El número de unidades de datos para el Agrupamiento 2 en T1 es de 310, y el número de unidades de datos para el Agrupamiento 2 en T0 es de 310. El número de unidades de datos para el Agrupamiento 3 en T1 es de 175, y el número de unidades de datos para el Agrupamiento 3 en T0 es de 165. El CDS 200a puede ver que el número de unidades de datos en los Agrupamientos 1 y 3 modificados de T0 a T1, mientras que el número de unidades de datos en el Agrupamiento 2 permaneció igual de T0 a T1. A partir de esta comparación, el CDS 200a puede determinar que los datos para el Agrupamiento 1 y el

Agrupamiento 3 pueden haber sido modificados y deberían obtenerse de la fuente 210 de datos. El CDS 200a puede hacer un seguimiento de los agrupamientos modificados 320, por ejemplo, para solicitar datos de estos agrupamientos de la fuente 210 de datos. Por ejemplo, la información de los agrupamientos modificados 320 puede enumerar los Agrupamientos 1 y 3.

El CDS 200a obtiene los datos para el Agrupamiento 1 de la fuente 210 de datos, y también obtiene los datos para el Agrupamiento 3 de la fuente 210 de datos (o cierto sumario de los agrupamientos, tales como los filtros de Bloom, en otras realizaciones). Se explicará adicionalmente el ejemplo con los datos obtenidos 330 del Agrupamiento 3. Los datos 330 del Agrupamiento 3 incluye la totalidad de las 175 unidades de datos incluidas en el agrupamiento. La fuente 210 de datos puede ser una base 210a de datos, y los datos 330 del Agrupamiento 3 pueden incluir filas como unidades de datos. Cada fila en los datos 330 del Agrupamiento 3 puede incluir la fecha y el momento para la fila (por ejemplo, el sello de tiempo de la columna actualizada por última vez) y los datos de esa fila.

El CDS 200a compara los datos 330 del Agrupamiento 3 con los datos históricos 335 del Agrupamiento 3. Los datos 330 del Agrupamiento 3 pueden asociarse con T1, y los datos históricos 335 del Agrupamiento 3 pueden asociarse con T0. Por ejemplo, los datos históricos 335 del Agrupamiento 3 pueden ser datos del Agrupamiento 3 que fueron obtenidos en T0. Los datos históricos del Agrupamiento 3 también pueden incluir filas como unidades de datos. Cada fila en los datos históricos 335 del Agrupamiento 3 también puede incluir la fecha y el momento para la fila y los datos de esa fila. Al comparar los datos 330 del Agrupamiento 3 y los datos históricos 335 del Agrupamiento 3, el CDS 200a puede determinar que la Fila 3 modificada, por ejemplo, la Fila 3 puede haber sido insertada después de T0. El CDS 200a marca la Fila 3 como una unidad de datos que ha de enviarse a un sistema 270 de destino. Al pasar por el resto de los datos 330 del Agrupamiento 3 y de los datos históricos 335 del Agrupamiento 3, el CDS 200a identifica 10 filas en este ejemplo que fueron añadidas. El CDS 200a puede hacer un seguimiento de las unidades modificadas de datos en una lista, tal como una lista 340 de unidades modificadas del Agrupamiento 3. En algunas realizaciones, en vez de comparar las unidades de datos con la anterior versión de las unidades de datos, el CDS 200a utiliza un filtro de Bloom al que se han añadido las unidades de datos en la anterior versión. El CDS 200a consulta el filtro de Bloom para determinar si una unidad de datos se encuentra en el conjunto.

El CDS 200a puede obtener información de agrupamiento para todas las fechas para las cuales hay disponible datos en la fuente 210 de datos. Por ejemplo, una fuente 210 de datos contiene datos de 1.000 días, el CDS 200a puede obtener la información de agrupamiento de la totalidad de los 1.000 días. O el CDS 200a puede especificar un marco temporal para el cual desea obtener información de agrupamiento, tal como 60 días. La información de agrupamiento puede obtenerse con facilidad de una fuente 210 de datos sin sobrecargar los recursos de la fuente 210 de datos. Al comparar la información histórica de agrupamiento, el CDS 200a puede identificar con facilidad qué agrupamientos pueden haber modificado datos.

Debido a que la comparación de la información de agrupamiento puede hacer que sea sencillo identificar fácilmente datos modificados durante un periodo prolongado de tiempo, el CDS 200a puede capturar todas las modificaciones en los datos. Por ejemplo, en un sistema que descarga datos para los últimos 5 días puede saltarse cualquier unidad de datos cuyo sello de tiempo actualizado por última vez haya sido modificado para encontrarse fuera de esta ventana de 5 días. Sin embargo, el CDS 200a puede detectar que se ha eliminado o añadido una unidad de datos a un agrupamiento particular en cualquier ventana de tiempo. Por ejemplo, un usuario modifica accidentalmente el sello de tiempo actualizado por última vez para la Fila 1 al Día 1 del Día 1.000. El sistema que solo descarga los últimos 5 días de datos se saltará la Fila 1, pero el CDS 200a reconocerá la Fila 1 como una modificación debido a que se reflejará en el número de unidades de datos para el Día 1 en la información de agrupamiento.

Se pueden seleccionar la unidad o el tamaño del agrupamiento y la columna de agrupamiento, de forma que la mayoría de los nuevos datos añadidos a la fuente 210 de datos se encuentre en una de las unidades de agrupamiento. En una realización, la unidad o el tamaño del agrupamiento puede estar relacionado con la latencia deseada del proceso, y la columna del agrupamiento puede relacionarse con la distribución. Se puede seleccionar la unidad o el tamaño del agrupamiento en distintos niveles de granularidad. Por ejemplo, un agrupamiento puede estar basado en una unidad de múltiples días, un día, múltiples horas, una hora, etc. Se puede seleccionar la unidad o el tamaño de un agrupamiento según sea apropiado, por ejemplo, en función de los requisitos de la fuente 210 de datos, del CDS 200a y/o del sistema 270 de destino. Se puede especificar la unidad o el tamaño de un agrupamiento a un nivel que proporciona una comparación significativa de agrupamientos. En algunas realizaciones, la unidad de agrupamiento que da lugar a una distribución homogénea de unidades de datos en agrupamientos puede no ser muy útil dado que cada agrupamiento tendrá una modificación al número de unidades de datos, y el CDS 200a tiene que comprobar casi todos los agrupamientos. Por ejemplo, si AGRUPAR POR era por una hora, en vez de un día, la partición para cada hora probablemente incluirá algunas filas, y casi todas las particiones tendrían que ser comprobadas, lo que puede dar lugar a obtener datos para la mayoría de las particiones. Por otra parte, AGRUPAR POR por un día dará lugar, probablemente, a datos añadidos recientemente que se encuentren en las particiones más recientes. Con un razonamiento similar, la columna o el campo utilizado para el agrupamiento por un criterio puede tener una característica que dé lugar a una distribución más “sesgada” que una distribución homogénea. En un ejemplo, si las unidades de datos fueron agrupadas por la primera letra del apellido de una persona, el agrupamiento para cada letra del alfabeto contendrá principalmente nuevas unidades de datos, y se tendrán que comprobar los agrupamientos para todas las letras del alfabeto. En otras realizaciones, puede desearse una

distribución homogénea de unidades de datos y, en consecuencia, se puede seleccionar la unidad o el tamaño del agrupamiento y la columna del agrupamiento para proporcionar una distribución homogénea de unidades de datos. Por ejemplo, esto puede hacerse de forma que no se coloque en un agrupamiento la mayoría de las unidades de datos que han de ser procesadas.

En el ejemplo en el que los datos son agrupados por la columna actualizada por última vez, el CDS 200s puede no diferenciar entre una unidad de datos que ha sido añadida y una unidad de datos existente que ha sido actualizada. En ciertas realizaciones, el CDS 200a puede implementar una forma de diferenciar entre los dos tipos de modificación. Por ejemplo, se puede asignar a cada unidad de datos un identificador único, por ejemplo, cuando se almacena la unidad de datos localmente en el almacenamiento local 250a. El identificador único puede ser utilizado para hacer un seguimiento de si se ha actualizado una unidad de datos. En este caso, es posible que el CDS 200a no pueda utilizar filtros de Bloom dado que no se almacenan los datos reales en filtros de Bloom.

En algunas realizaciones, el CDS 200a puede reconocer que se han borrado algunas unidades de datos. Por ejemplo, se puede reducir el número de unidades de datos para un agrupamiento en comparación con el anterior número de unidades de datos para ese agrupamiento. El CDS 200a puede identificar los datos borrados comparando los datos para el agrupamiento con la versión local de los datos para el agrupamiento. El CDS 200a puede enviar información al sistema 270 de destino de que se han borrado las unidades identificadas de datos, y el sistema 270 de destino puede borrar las unidades de datos de su almacenamiento en función de la información enviada por el CDS 200a.

Según se ha descrito anteriormente, el CDS 200a puede ofrecer muchas ventajas. El CDS 200a puede identificar un subconjunto modificado de datos en una fuente 210 de datos sin descargar todos los datos. El CDS 200a puede hacerlo para grandes cantidades, lo que puede ser muy eficaz. Solo se descarga una porción de los datos que puede incluir modificaciones para extraer la modificación real. El CDS 200a también puede identificar modificaciones de una forma genérica y puede funcionar con diversas fuentes 210 de datos. A menudo, el CDS 200a puede no tener cualquier información acerca de los datos de una fuente 210 de datos. Por ejemplo, el CDS 200a puede no saber cómo se estructuran los datos (por ejemplo, esquema de la base de datos, formato del fichero, etc.), cuán frecuentemente se actualizan los datos, formas en las que se actualizan los datos, o cómo se actualizan los datos. El CDS 200a puede identificar una columna, tal como la columna actualizada por última vez que puede indicar si una unidad de datos podría ser nueva y proceder a identificar modificaciones llevando a cabo un agrupamiento por un criterio en la columna seleccionada. El CDS 200a también puede gestionar datos en distintos formatos, tales como bases de datos y ficheros, de la misma forma, o similar. Debido a que el CDS 200a solo obtiene o capta datos modificados de una fuente 210 de datos para enviarlos a un sistema 270 de destino, el CDS 200a también puede ser denominado “captador”.

La Figura 4A es un diagrama de flujo que ilustra una realización de un procedimiento 400a para determinar y obtener las modificaciones en los datos de una pluralidad de fuentes de datos. El procedimiento 400a puede ser implementado por uno o más sistemas descritos con respecto a las Figuras 1-2 y 5. Con fines ilustrativos, se explica a continuación el procedimiento 400a en conexión con el CDS 200a en la Figura 2A. Con respecto a las Figuras 1-5 se explican con más detalle ciertos detalles relativos al procedimiento 400a. Dependiendo de la realización, el procedimiento 400a puede incluir menos bloques, o adicionales, y se pueden llevar a cabo los bloques en un orden que es distinto de lo ilustrado.

En el bloque 401a, el CDS 200a obtiene información que indica agrupamientos de datos de una fuente 210 de datos. Los datos pueden ser almacenados en uno o más ficheros o bases de datos en la fuente 210 de datos. La información puede indicar un número de unidades de datos incluidas en cada uno de los agrupamientos. Los agrupamientos pueden estar basados en sellos de tiempo de unidades respectivas de datos. Los sellos de tiempo pueden indicar momentos respectivos en los que se actualizaron por última vez las unidades de datos. En ciertas realizaciones, los sellos de tiempo de las unidades respectivas de datos incluyen la fecha y el momento en el que se actualizaron por última vez las unidades respectivas de datos, y los agrupamientos se basan únicamente en la fecha de los sellos de tiempo de las unidades respectivas de datos. Por ejemplo, se lleva a cabo una operación de agrupamiento únicamente en función de la fecha de los sellos de tiempo asociados con las unidades de datos. En tal caso, cada agrupamiento se asocia con una fecha específica. En algunas realizaciones, los agrupamientos se basan en el campo de unidades respectivas de datos que pueden proporcionar una distribución no homogénea de unidades de datos incluidas en cada agrupamiento.

El CDS 200a puede obtener la información que indica los agrupamientos de los datos en un intervalo. El CDS 200a también puede obtener la información que indica los agrupamientos de los datos en respuesta a la recepción de una solicitud procedente de un sistema 270 de destino. El CDS 200a puede obtener información que indica los agrupamientos de los datos almacenados en uno o más ficheros en la fuente 210 de datos utilizando un primer adaptador. El CDS 200a puede obtener información que indica los agrupamientos de los datos almacenados en una o más bases de datos en la fuente 210 de datos utilizando un segundo adaptador. El primer adaptador y el segundo adaptador pueden ser distintos.

En el bloque 402a, el CDS 200a determina un agrupamiento cuyas unidades de datos han sido modificadas. El CDS 200a puede determinar si las unidades de datos de un agrupamiento han sido modificadas al comparar un número de unidades de datos incluidas en el agrupamiento y un número histórico de unidades de datos incluidas en una versión local correspondiente de ese agrupamiento. La versión local correspondiente del primer agrupamiento puede ser creada en función de las unidades de datos incluidas en el agrupamiento en un momento anterior a la obtención de la información que indica los agrupamientos de los datos. Se puede denominar a este momento instante T0.

En el bloque 403a, el CDS 200a obtiene unidades de datos en el agrupamiento modificado procedentes de la fuente 210 de datos. Una unidad de datos incluida en el agrupamiento puede ser una fila en las una o más bases de datos de la fuente 210 de datos o una línea en los uno o más ficheros en la fuente 210 de datos. En ciertas realizaciones, si la fuente 210 de datos incluye uno o más ficheros, el CDS 200a puede comprobar el sello de tiempo de un fichero y compararlo con el sello de tiempo de la versión anterior del fichero para determinar si el fichero puede incluir nuevos datos. Al comparar los sellos de tiempo del fichero actual y de la versión anterior del fichero, el CDS 200a no necesita analizar los datos en el fichero para determinar si se han añadido nuevos datos. En tales realizaciones, el CDS 200a puede no obtener información de agrupamiento. Además, con respecto a los bloques 402a y 403a, el CDS 200a puede comparar directamente las unidades de datos en el fichero actual y las unidades de datos en la versión anterior del fichero, en vez de determinar uno o más agrupamientos modificados y/u obtener las unidades de datos en el agrupamiento modificado de la fuente 210 de datos.

En el bloque 404a, el CDS 200a compara las unidades de datos en el agrupamiento con las unidades de datos de la versión local correspondiente del agrupamiento. Al comparar las unidades de datos, el CDS 200 puede determinar qué unidades de datos del agrupamiento de la fuente 210 de datos han sido modificadas. La versión local correspondiente del agrupamiento puede incluir una copia de las unidades de datos incluidas en el agrupamiento en T0. En ciertas realizaciones, cuando la fuente 210 de datos incluye uno o más ficheros, el CDS 200a puede tratar cada nueva línea en el fichero como una unidad de datos y comparar las unidades de datos en el fichero actual y las unidades de datos en la versión anterior del fichero para identificar las unidades modificadas de datos.

En el bloque 405a, el CDS 200a extrae las unidades modificadas de datos del agrupamiento. El CDS 200a puede remitir las unidades modificadas extraídas de datos a uno o más sistemas 270 de destino.

Si el número de unidades de datos incluidas en el agrupamiento es superior al número histórico de unidades de datos incluidas en la versión local correspondiente del agrupamiento, el CDS 200a puede identificar las unidades modificadas de datos como unidades añadidas o actualizadas de datos, y remitir las unidades modificadas de datos al sistema 270 de destino para ser almacenadas. Si el número de unidades de datos incluidas en el agrupamiento es menor que el número histórico de unidades de datos incluidas en la versión local correspondiente del agrupamiento, el CDS 200a puede identificar las unidades modificadas de datos como unidades borradas de datos, y remitir las unidades modificadas de datos al sistema 270 de destino para que sean eliminadas.

En algunas realizaciones, el CDS 200a asigna un identificador único a cada una de las unidades de datos incluidas en el agrupamiento. El CDS 200a puede determinar si una unidad modificada es una nueva unidad de datos o una unidad actualizada de datos en función del identificador único asociado con la unidad modificada de datos.

La Figura 4B es un diagrama de flujo que ilustra otra realización de un procedimiento 400b para determinar y obtener las modificaciones en los datos de una pluralidad de fuentes de datos. El procedimiento 400b puede implementarse mediante uno o más sistemas descritos con respecto a las Figuras 1-2 y 5. Con fines ilustrativos, se explica a continuación el procedimiento 400b en conexión con el CDS 200b en la Figura 2B. Se explican con más detalle con respecto a las Figuras 1-5 ciertos detalles relativos al procedimiento 400b. Dependiendo de la realización, el procedimiento 400b puede incluir menos bloques, o adicionales, y los bloques pueden llevarse a cabo en un orden distinto al ilustrado.

En el bloque 401b, el CDS 200b obtiene información que indica los agrupamientos de datos de una fuente 210 de datos. La fuente 210 de datos puede ser una base de datos o un fichero. La información puede indicar un número de unidades de datos incluidas en cada uno de los agrupamientos. Los agrupamientos pueden estar basados en los sellos de tiempo de unidades respectivas de datos. Los sellos de tiempo pueden indicar momentos respectivos en los que se actualizaron por última vez las unidades de datos.

En el bloque 402b, el CDS 200b determina un agrupamiento cuyas unidades de datos han sido modificadas. El CDS 200b puede determinar si las unidades de datos de un agrupamiento han sido modificadas comparando un número de unidades de datos incluidas en el agrupamiento y un número histórico de unidades de datos incluidas en cada uno de los agrupamientos. Por ejemplo, el número histórico de unidades de datos incluidas en cada uno de los agrupamientos puede ser almacenado en datos históricos 315 de agrupamiento expuestos con respecto a la Figura 3.

En el bloque 403b, el CDS 200b obtiene unidades de datos en el agrupamiento modificado procedente de la fuente 210 de datos. Si la fuente 210 de datos es una base 210a de datos, una unidad de datos incluida en el agrupamiento puede ser una fila, y si la fuente 210 de datos es un fichero 210b, una unidad de datos incluida en el agrupamiento puede ser una línea.

En el bloque 404b, el CDS 200b compara las unidades de datos en el agrupamiento con una versión comprimida de los datos. El CDS 200b puede comparar las unidades de datos en el agrupamiento con una versión local correspondiente del agrupamiento, que puede ser una versión comprimida de los datos. En función de la comparación, el CDS 200b puede determinar qué unidades de datos del agrupamiento procedente de la fuente 210 de datos han sido modificadas. La versión comprimida de los datos puede ser una estructura probabilística de datos que ahorra espacio, tal como un filtro de Bloom. La estructura probabilística de datos que ahorra espacio puede incluir información acerca de unidades de datos incluidas en el agrupamiento en un momento anterior a la obtención de la información que indica los agrupamientos de los datos de la fuente 210 de datos. Este momento puede ser denominado instante T0. La estructura probabilística de datos que ahorra espacio puede identificar si se incluyó una unidad de datos incluida en el agrupamiento en el momento anterior (por ejemplo, el instante T0). En algunas realizaciones, la estructura probabilística de datos que ahorra espacio es un filtro de Bloom. Se puede seleccionar el filtro de Bloom entre múltiples filtros de Bloom, teniendo cada uno un tamaño distinto. La versión comprimida de los datos puede no comprender una copia del agrupamiento.

La versión comprimida de los datos se almacena en el almacenamiento local 250b, y las unidades modificadas de datos extraídas remitidas al sistema 270 de destino son almacenadas en el almacenamiento en el sistema 270 de destino. En ciertas realizaciones, el almacenamiento local 250b tiene una menor capacidad de almacenamiento que el almacenamiento del sistema 270 de destino. En una realización, el almacenamiento local 250b incluye un NAS.

En el bloque 405b, el CDS 200b extrae las unidades modificadas de datos del agrupamiento. En el bloque 406b, el CDS 200b remite las unidades modificadas de datos al sistema 270 de destino. Los bloques 405b y 406b pueden ser similares a los bloques 405a en la Figura 4A.

Mecanismos de implementación

Según una realización, las técnicas descritas en la presente memoria son implementadas por uno o más dispositivos informáticos de uso especial. Los dispositivos informáticos de uso especial pueden ser cableados para llevar a cabo las técnicas, o pueden incluir circuitería o dispositivos electrónicos digitales tales como uno o más circuitos integrados para aplicaciones específicas (ASIC) o matrices de puertas de campo programable (FPGA) que son programados de manera persistente para llevar a cabo las técnicas, o pueden incluir uno o más procesadores de soporte físico programados para llevar a cabo las técnicas de acuerdo con instrucciones del programa en soporte lógico inalterable, en memoria, en otro almacenamiento, o una combinación. Tales dispositivos informáticos de uso especial también pueden combinar lógica cableada a medida, ASIC, FPGA con programación a medida para lograr las técnicas. Los dispositivos informáticos de uso especial pueden ser sistemas de ordenador de sobremesa, sistemas de ordenador servidor, sistemas de ordenador portátil, dispositivos portátiles, dispositivos de red o cualquier otro dispositivo o combinación de dispositivo que incorpore soporte físico y/o soporte lógico para implementar las técnicas.

El o los dispositivos informáticos están controlados y coordinados, en general, mediante soporte lógico de sistema operativo, tal como iOS, Android, Chrome OS, Windows XP, Windows Vista, Windows 7, Windows 8, Windows Server, Windows CE, Unix, Linux, SunOS, Solaris, iOS, Blackberry OS, Vx-Works u otros sistemas operativos compatibles. En otras realizaciones, se puede controlar el dispositivo informático mediante un sistema operativo patentado. Los sistemas operativos convencionales controlan y planifican procedimientos de ordenador para su ejecución, llevar a cabo una gestión de la memoria, proporcionar un sistema de ficheros, conexión de red, servicios de I/O y proporcionar una funcionalidad de interfaz de usuario, tal como una interfaz gráfica de usuario (“GUI”), entre otros.

Por ejemplo, la Figura 8 es un diagrama de bloques que ilustra un sistema 500 de ordenador en el que se puede implementar una realización. Por ejemplo, el sistema informático 500 puede comprender un sistema servidor que accede a datos de cuerpos policiales y proporciona datos de interfaz de usuario a uno o más usuarios (por ejemplo, administradores) que permite a esos usuarios visualizar sus paneles de control administrativos deseados e interactuar con los datos. Otros sistemas informáticos divulgados en la presente memoria, tales como el usuario (por ejemplo, administrador), puede incluir cualquier porción de la circuitería y/o la funcionalidad expuesta con referencia al sistema 500.

El sistema 500 de ordenador incluye un *bus* 502 u otro mecanismo de comunicación para comunicar información, y un procesador, o múltiples procesadores, 504 de soporte físico acoplado con el *bus* 502 para procesar información. El o los procesadores 504 de soporte físico pueden ser, por ejemplo, uno o más microprocesadores de uso general.

El sistema 500 de ordenador también incluye una memoria principal 506, tal como una memoria de acceso aleatorio (RAM), una memoria intermedia y/u otros dispositivos de almacenamiento dinámico, acoplada con el *bus* 502 para almacenar información e instrucciones que han de ser ejecutadas por el procesador 504. La memoria principal 506 también puede ser utilizada para almacenar variables temporales u otra información intermedia durante la ejecución de instrucciones que han de ser ejecutadas por el procesador 504. Tales instrucciones, cuando son almacenadas en medios de almacenamiento accesibles al procesador 504, convierten al sistema 500 de ordenador en una máquina de uso especial que se personaliza para llevar a cabo las operaciones especificadas en las instrucciones.

El sistema 500 de ordenador incluye, además, una memoria 808 de solo lectura (ROM) u otro dispositivo de almacenamiento estático acoplado con el *bus* 502 para almacenar información estática e instrucciones para el procesador 504. Se proporciona un dispositivo 510 de almacenamiento, tal como un disco magnético, disco óptico, o unidad de lápiz de USB (unidad *flash*), etc., y se acopla con el *bus* 502 para almacenar información e instrucciones.

El sistema 500 de ordenador puede estar acoplado mediante el *bus* 502 con un medio 512 de visualización, tal como un tubo de rayos catódicos (CRT) o pantalla LCD (o pantalla táctil), para representar visualmente información a un usuario del ordenador. Un dispositivo 514 de entrada, que incluye teclas alfanuméricas y otras, está acoplado con el *bus* 502 para comunicar información y selecciones de instrucciones al procesador 504. Otro tipo de dispositivo de entrada de usuario es un control 516 del cursor, tal como un ratón, una bola de desplazamiento o teclas de dirección del cursor para comunicar información de dirección y selecciones de instrucciones al procesador 504 y para controlar el movimiento del cursor en el medio 512 de visualización. Normalmente, el dispositivo de entrada tiene dos grados de libertad en dos ejes, un primer eje (por ejemplo, x) y un segundo eje (por ejemplo, y), lo que permite al dispositivo especificar las posiciones en un plano. En algunas realizaciones, se puede implementar la misma información de dirección y las selecciones de instrucciones como un control del cursor mediante la recepción de toques en una pantalla táctil sin un cursor.

El sistema informático 500 puede incluir un módulo de interfaz de usuario para implementar una GUI que puede ser almacenada en un dispositivo de almacenamiento masivo como códigos ejecutables de soporte lógico que son ejecutados por el o los dispositivos informáticos. Este y otros módulos pueden incluir, a modo de ejemplo, componentes, tales como componentes de soporte lógico, componentes de soporte lógico orientados a objetos, componentes de clases y componentes de tareas, procesos, funciones, atributos, procedimientos, subrutinas, segmentos de código de programa, controladores, soporte lógico inalterable, microcódigo, circuitería, datos, bases de datos, estructuras de datos, tablas, matrices y variables.

En general, según se utiliza en la presente memoria, la palabra “módulo”, hace referencia a lógica implementada en soporte físico o en soporte lógico inalterable, o a una colección de instrucciones de soporte lógico, que tienen, posiblemente, puntos de entrada y de salida, escritas en un lenguaje de programación, tal como, por ejemplo, Java, Lua, C o C++. Se puede compilar y enlazar un módulo de soporte lógico en un programa ejecutable, instalado en una biblioteca de enlace dinámico, o puede escribirse en un lenguaje interpretado de programación tal como, por ejemplo, BASIC, Perl o Python. Se apreciará que los módulos de soporte lógico pueden ser susceptibles de ser llamados desde otros módulos o desde ellos mismos, y/o pueden ser invocados en respuesta a eventos o interrupciones detectados. Los módulos de soporte lógico configurados para su ejecución en dispositivos informáticos pueden proporcionarse en un medio legible por un ordenador, tal como un disco compacto, un disco de vídeo digital, una unidad *flash*, un disco magnético, o cualquier otro medio tangible, o como una descarga digital (y puede ser almacenada originalmente en un formato comprimido o instalable que requiere su instalación, descompresión o decodificación antes de su ejecución). Tal código de soporte lógico puede ser almacenado, parcial o completamente, en un dispositivo de memoria del dispositivo informático ejecutante, para su ejecución por el dispositivo informático. Las instrucciones de soporte lógico pueden ser embebidas en el soporte lógico inalterable, tal como una EPROM. Se apreciará, además, que los módulos de soporte físico pueden comprender unidades lógicas conectadas, tales como puertas y circuitos eléctricos biestables, y/o pueden comprender unidades programables, tales como procesadores o matrices de puertas programables. La funcionalidad de los módulos o del dispositivo informático descrita anteriormente en la presente memoria puede implementarse, preferentemente, como módulos de soporte lógico, pero puede representarse en soporte físico o en soporte lógico inalterable. En general, los módulos descritos en la presente memoria hacen referencia a módulos lógicos que pueden combinarse con otros módulos o ser divididos en submódulos a pesar de su organización física o almacenamiento.

El sistema 500 de ordenador puede implementar las técnicas descritas en la presente memoria utilizando soporte físico a medida, uno o más ASIC o FPGA, soporte lógico inalterable y/o soporte lógico que, en combinación con el sistema de ordenador, provoca o programa el sistema 500 de ordenador para que sea una máquina de uso especial. Según una realización, las técnicas se llevan a cabo en la presente memoria mediante el sistema 500 de ordenador en respuesta al o a los procesadores 504 que ejecutan una o más secuencias de una o más instrucciones contenidas en la memoria principal 506. Tales instrucciones pueden ser leídas en la memoria principal 506 desde otro medio de almacenamiento, tal como el dispositivo 510 de almacenamiento. La ejecución de las secuencias de instrucciones contenidas en la memoria principal 506 provoca que el o los procesadores 504 lleven a cabo las etapas del procedimiento descritas en la presente memoria. En realizaciones alternativas, se puede utilizar circuitería cableada en vez de instrucciones de soporte lógico, o en combinación con las mismas.

Según se utiliza en la presente memoria, la expresión “medios no transitorios” y expresiones similares hacen referencia a cualquier medio que almacene datos y/o instrucciones que provocan que una máquina opere de una forma específica. Tales medios no transitorios pueden comprender medios no volátiles y/o medios volátiles. Los medios no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, tales como el dispositivo 510 de almacenamiento. Los medios volátiles incluyen memoria dinámica, tal como la memoria principal 506. Formas comunes de medios no transitorios incluyen, por ejemplo, un disquete, un disco duro, una unidad de estado sólido, una cinta magnética o cualquier otro medio de almacenamiento magnético de datos, un CD-ROM, cualquier otro medio de almacenamiento óptico de datos, cualquier medio físico con patrones de agujeros, una RAM, una PROM,

una EPROM, una FLASH-EPROM, una NVRAM, cualquier otro *chip* o cartucho de memoria, y versiones en red de los mismos.

Los medios no transitorios son distintos de los medios de transmisión, pero pueden ser utilizados junto con los mismos. Los medios de transmisión participan transfiriendo información entre medios no transitorios. Por ejemplo, los medios de transmisión incluyen cables coaxiales, hilo de cobre y fibra óptica, incluyendo los hilos que comprenden el *bus* 502. Los medios de transmisión también pueden adoptar la forma de ondas acústicas o de luz, tales como las generadas durante comunicaciones de datos por ondas de radio y por infrarrojos.

Diversas formas de medios pueden estar implicadas en la realización de una o más secuencias de una o más instrucciones al procesador 504 para su ejecución. Por ejemplo, las instrucciones pueden ser portadas, inicialmente, en un disco magnético o en una unidad de estado sólido de un ordenador remoto. El ordenador remoto puede cargar las instrucciones en su memoria dinámica y enviar las instrucciones por una línea telefónica utilizando un módem. Un módem local al sistema 500 de ordenador puede recibir los datos en la línea telefónica y utilizar un transmisor de rayos infrarrojos para convertir los datos en una señal de rayos infrarrojos. Un detector de rayos infrarrojos puede recibir los datos portados en la señal infrarroja y una circuitería apropiada puede situar los datos en el *bus* 502. El *bus* 502 transporta los datos hasta la memoria principal 506, desde la que el procesador 504 recupera y ejecuta las instrucciones. Las instrucciones recibidas por la memoria principal 506 puede recuperar y ejecutar las instrucciones. Las instrucciones recibidas por la memoria principal 506 pueden ser almacenadas, opcionalmente, en el dispositivo 510 de almacenamiento, bien antes o bien después de su ejecución por el procesador 504.

El sistema 500 de ordenador también incluye una interfaz 518 de comunicaciones acoplada con el *bus* 502. La interfaz 518 de comunicaciones proporciona un acoplamiento bidireccional de comunicaciones de datos con un enlace 520 de red que está conectado con una red local 522. Por ejemplo, la interfaz 518 de comunicaciones puede ser una tarjeta de red digital de servicios integrados (ISDN), un módem de cable, un módem por satélite, o un módem para proporcionar una conexión de comunicaciones de datos con un tipo correspondiente de línea telefónica. Como otro ejemplo, la interfaz 518 de comunicaciones puede ser una tarjeta de una red de área local (LAN) para proporcionar una conexión de comunicaciones de datos con una LAN compatible (o componente de WAN que ha de comunicarse con una WAN). También se pueden implementar enlaces inalámbricos. En cualquier implementación tal, la interfaz 518 de comunicaciones envía y recibe señales eléctricas, electromagnéticas u ópticas que transportan corrientes de datos digitales que representan diversos tipos de información.

Normalmente, el enlace 520 de red proporciona una comunicación de datos a través de una o más redes a otros servicios de datos. Por ejemplo, el enlace 520 de red puede proporcionar una conexión a través de la red local 522 con un ordenador anfitrión 524 o con equipos de datos operados por un proveedor de servicios de Internet (ISP) 526. A su vez, el ISP 526 proporciona servicios de comunicaciones de datos a través de la red mundial de comunicaciones de paquetes de datos denominada ahora, habitualmente, “Internet” 525. Tanto la red local 522 como Internet 525 utilizan señales eléctricas, electromagnéticas u ópticas que transportan corrientes de datos digitales. Las señales a través de las diversas redes y las señales en el enlace 520 de red y a través de la interfaz 518 de comunicaciones, que transporta los datos digitales al sistema 500 de ordenador, y desde el mismo, son formas ejemplares de medios de transmisión.

El sistema 500 de ordenador puede enviar mensajes y recibir datos, incluyendo código de programa, a través de la o las redes, del enlace 520 de red y de la interfaz 518 de comunicaciones. En el ejemplo de Internet, un servidor 530 podría transmitir un código solicitado para un programa de aplicación a través de Internet 525, del ISP 526, de la red local 522 y de la interfaz 518 de comunicaciones.

El código recibido puede ser ejecutado por el procesador 504 como es recibido y/o almacenado en el dispositivo 510 de almacenamiento, u otro almacenamiento no volátil para una ejecución posterior.

Cada uno de los procedimientos, procedimientos y algoritmos descritos en las secciones precedentes puede ser implementado en módulos de código, y automatizado completa o parcialmente en los mismos, ejecutado por uno o más sistemas de ordenador o procesadores de ordenador que comprenden soporte físico de ordenador. Los procedimientos y algoritmos pueden ser implementados parcial o completamente en circuitería para aplicaciones específicas.

Los diversos procedimientos y características descritos anteriormente pueden ser utilizados con independencia mutua, o pueden combinarse de diversas formas. Se concibe que todas las posibles combinaciones y subcombinaciones se encuentren dentro del alcance de la presente divulgación. Además, se pueden omitir en algunas implementaciones ciertos bloques de procedimiento o de proceso. Los procedimientos y procesos descritos en la presente memoria tampoco están limitados a ninguna secuencia particular, y los bloques o estados relativos a la misma pueden llevarse a cabo en otras secuencias que sean apropiadas. Por ejemplo, los bloques o estados descritos pueden llevarse a cabo en un orden distinto del divulgado específicamente, o se pueden combinar múltiples bloques o estados en un único bloque o estado. Los bloques o estados ejemplares pueden llevarse a cabo en serie, en paralelo o de alguna otra forma. Se pueden añadir bloques o estados a las realizaciones ejemplares divulgadas, o pueden ser eliminados de las mismas. Los sistemas y componentes ejemplares descritos en la presente memoria pueden estar configurados de forma distinta a lo descrito. Por ejemplo, se pueden añadir

elementos a las realizaciones ejemplares divulgadas, o pueden ser eliminados de las mismas o reordenados en comparación con las mismas.

El lenguaje condicional, tal como, entre otros, “puede”, “podría”, “pudiera” o “puede que”, a no ser que se indique específicamente lo contrario, o se comprenda lo contrario en el contexto usado, se concibe, en general, que transmita que ciertas realizaciones incluyen, ciertos elementos, características y/o etapas, mientras que otras realizaciones no lo hacen. Por lo tanto, no se concibe, en general, que no se requiera de ninguna manera que tal lenguaje condicional implique que los elementos, características y/o etapas para una o más realizaciones o que una o más realizaciones incluyan necesariamente lógica para decidir, con o sin una entrada o indicación por parte del usuario, si se incluyen estos elementos, características y/o etapas o deben ser llevados a cabo en cualquier realización particular.

Se debería comprender que cualquier descripción, elemento o bloque del procedimiento en los diagramas de flujo descritos en la presente memoria y/o mostrados en las figuras adjuntas representa potencialmente módulos, segmentos o porciones de código que incluyen una o más instrucciones ejecutables para implementar funciones o etapas lógicas específicas en el procedimiento. Se incluyen implementaciones alternativas en el alcance de las realizaciones descritas en la presente memoria en las que se pueden borrar, ejecutar fuera del orden mostrado o expuesto o presentado elementos o funciones, incluyendo de forma sustancialmente simultánea o en un orden inverso, dependiendo de la funcionalidad implicada, según comprenderán los expertos en la técnica.

Diversos aspectos de la presente invención pueden estar relacionados con una o más de las siguientes realizaciones ejemplares enumeradas (EEE), cada una de las cuales son ejemplos. Como con cualquier otra exposición relacionada proporcionada en el presente documento, no se debería interpretar que las EEE limitan cualquier reivindicación.

EEE1: Un sistema de ordenador configurado para obtener datos modificados de una fuente de datos, comprendiendo el sistema de ordenador: uno o más procesadores de soporte físico de ordenador configurados para ejecutar código para provocar que el sistema: obtenga información que indica una pluralidad de agrupamientos de datos almacenados en uno o más ficheros o bases de datos en una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determine un primer agrupamiento de la pluralidad de agrupamientos que incluyen una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un primer número histórico de unidades de datos incluidas en una versión local correspondiente del primer agrupamiento, en el que se crea la versión local correspondiente del primer agrupamiento en función de las unidades de datos incluidas en el primer agrupamiento en un primer momento anterior a dicha obtención de la información que indica la pluralidad de agrupamientos de los datos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de la versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

EEE2: El sistema de la EEE 1, en el que la pluralidad de agrupamientos está basada en sellos de tiempo de unidades respectivas de datos, indicando los sellos de tiempo los momentos respectivos en los que se actualizaron por última vez las unidades de datos.

EEE 3: El sistema de la EEE 1 o EEE 2, en el que los sellos de tiempo de las unidades respectivas de datos incluyen una fecha y un momento en los que se actualizaron por última vez las unidades respectivas de datos, y en el que la pluralidad de agrupamientos se basa únicamente en la fecha de los sellos de tiempo de las unidades respectivas de datos.

EEE 4: El sistema de la EEE 3, en el que cada uno de la pluralidad de agrupamientos está asociado con una fecha distinta.

EEE 5: El sistema de la EEE 3 o EEE 4, en el que las fechas asociadas con cada uno de la pluralidad de agrupamientos se encuentran en un intervalo de fechas.

EEE 6: El sistema de cualquiera de las EEE 1-5, en el que la pluralidad de agrupamientos se basa en un primer campo de unidades respectivas de datos, configurado el primer campo para proporcionar una distribución no homogénea de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos.

EEE 7: El sistema de cualquiera de las EEE 1-6, en el que la versión local correspondiente del primer agrupamiento comprende una copia de las unidades de datos incluidas en el primer agrupamiento en el primer momento.

EEE 8: El sistema de cualquiera de las EEE 1-7, en el que, en respuesta a la determinación de que el primer número de unidades de datos es mayor que el primer número histórico de unidades de datos, el código está configurado,

además, para provocar que el sistema: identifique las unidades modificadas de datos como unidades añadidas o modificadas de datos; y remita las unidades modificadas de datos al sistema de destino para ser almacenadas.

EEE 9: El sistema de cualquiera de las EEE 1-8, en el que, en respuesta a la determinación de que el primer número de unidades de datos es menor que el primer número histórico de unidades de datos, el código está configurado, además, para provocar que el sistema: identifique las unidades modificadas de datos como unidades borradas de datos; y remita las unidades modificadas de datos al sistema de destino para ser eliminadas.

EEE 10: El sistema de cualquiera de las EEE 1-9, en el que el código está configurado adicionalmente para: asignar un identificador único a cada una de las unidades de datos incluidas en el primer agrupamiento; y determinar si una primera unidad modificada de datos de las unidades modificadas de datos es una nueva unidad de datos o una unidad actualizada de datoss en función del identificador único asociado con la primera unidad modificada de datos.

EEE 11: El sistema de cualquiera de las EEE 1-10, en el que una unidad de datos incluida en el primer agrupamiento es una fila en una o más bases de datos o una línea en los uno o más ficheros.

EEE 12: El sistema de cualquiera de las EEE 1-11, en el que el código está configurado, además, para provocar que el sistema: obtenga la información que indica la pluralidad de agrupamientos de los datos almacenados en los uno o más ficheros en la fuente de datos utilizando un primer adaptador; y obtenga la información que indica la pluralidad de agrupamientos de los datos almacenados en las una o más bases de datos en la fuente de datos utilizando un segundo adaptador.

EEE 13: El sistema de cualquiera de las EEE 1-12, en el que el código está configurado, además, para provocar que el sistema obtenga la información que indica la pluralidad de agrupamientos de los datos en un intervalo.

EEE 14: El sistema de cualquiera de las EEE 1-13, en el que el código está configurado, además, para provocar que el sistema obtenga la información que indica la pluralidad de agrupamientos de los datos en respuesta a la recepción de una solicitud procedente del sistema de destino.

EEE 15: Un procedimiento para obtener datos modificados de una fuente de datos, comprendiendo el procedimiento: obtener, mediante uno o más procesadores de soporte físico de ordenador, información que indica una pluralidad de agrupamientos de datos almacenados en uno o más ficheros o bases de datos en una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determinar, mediante los uno o más procesadores de soporte físico de ordenador, un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un primer número histórico de unidades de datos incluidas en una versión local correspondiente del primer agrupamiento, en el que la versión local correspondiente del primer agrupamiento se crea en función de unidades de datos incluidas en el primer agrupamiento en un primer momento antes de dicha obtención de la información que indica la pluralidad de agrupamientos de los datos; acceder, mediante los uno o más procesadores de soporte físico de ordenador, unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; comparar, mediante los uno o más procesadores de soporte físico de ordenador, las unidades de datos incluidas en el primer agrupamiento con unidades de datos de la versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas; extraer, mediante los uno o más procesadores de soporte físico de ordenador, las unidades modificadas de datos del primer agrupamiento; y remitir, mediante los uno o más procesadores de soporte físico de ordenador, las unidades modificadas extraídas de datos a un sistema de destino.

EEE 16: El procedimiento de la EEE 15, en el que la pluralidad de agrupamientos se basa en sellos de tiempo de unidades respectivas de datos, indicando los sellos de tiempo momentos respectivos en los que se actualizaron por última vez las unidades de datos.

EEE 17: El procedimiento de la EEE 15 o de la EEE 16, en el que la pluralidad de agrupamientos se basa en un primer campo de unidades respectivas de datos, el primer campo configurado para proporcionar una distribución no homogénea de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos.

EEE 18: El procedimiento de cualquiera de las EEE 15-17, en el que la versión local correspondiente del primer agrupamiento comprende una copia de las unidades de datos incluidas en el primer agrupamiento en el primer momento.

EEE 19: El procedimiento de cualquiera de las EEE 15-18, en el que una unidad de datos incluida en el primer agrupamiento es una fila en las una o más bases de datos o una línea en los uno o más ficheros.

EEE 20: Un medio transitorio o no transitorio legible por un ordenador que comprende instrucciones para obtener datos modificados de una fuente de datos que provoca que un procesador de ordenador: obtenga información que indica una pluralidad de agrupamientos de datos almacenados en uno o más ficheros o bases de datos en una

fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determine un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un primer número histórico de unidades de datos incluidas en una versión local correspondiente del primer agrupamiento, creándose la versión local correspondiente del primer agrupamiento en función de las unidades de datos incluidas en el primer agrupamiento en un primer momento antes de dicha obtención de la información que indica la pluralidad de agrupamientos de los datos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de la versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

EEE 21: Un sistema de ordenador configurado para obtener datos modificados de una fuente de datos, comprendiendo el sistema de ordenador: uno o más procesadores de soporte físico de ordenador configurados para ejecutar código para provocar que el sistema: obtenga información que indica una pluralidad de agrupamientos de datos de una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determine un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un número histórico de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de una versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas, en el que la versión local correspondiente del primer agrupamiento de unidades de datos es una versión comprimida del primer agrupamiento de unidades de datos; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

EEE 22: El sistema de la EEE 21, en el que la versión comprimida del primer agrupamiento de unidades de datos es una estructura probabilística de datos que ahorra espacio que incluye información acerca de las unidades de datos incluidas en el primer agrupamiento en un primer momento antes de dicha obtención de la información que indica la pluralidad de agrupamientos de los datos de la fuente de datos.

EEE 23: El sistema de la EEE 22, en el que la estructura probabilística de datos que ahorra espacio está configurada para determinar si una unidad particular de datos incluida en el primer agrupamiento fue incluida en el primer agrupamiento en el primer momento.

EEE 24: El sistema de la EEE 22 o de la EEE 23, en el que la estructura probabilística de datos que ahorra espacio es un filtro de Bloom.

EEE 25: El sistema de la EEE 23 o de la EEE 24, en el que el filtro de Bloom se selecciona de una pluralidad de filtros de Bloom cada uno de los cuales puede incluir un número distinto de unidades de datos.

EEE 26: El sistema de cualquiera de las EEE 21-25, en el que la versión comprimida de los datos no comprende una copia de las unidades de datos del primer agrupamiento.

EEE 27: El sistema de cualquiera de las EEE 21-26, en el que la versión local correspondiente de los datos se almacena en un primer almacenamiento y las unidades modificadas extraídas de datos remitidas al sistema de destino son almacenadas en un segundo almacenamiento, teniendo el primer almacenamiento una menor capacidad de almacenamiento que el segundo almacenamiento.

EEE 28: El sistema de la EEE 27, en el que el primer almacenamiento es un almacenamiento conectado en red (NAS).

EEE 29: El sistema de cualquiera de las EEE 21-28, en el que la fuente de datos es una base de datos o un fichero.
EEE 30: El sistema de cualquiera de las EEE 21-29, en el que la pluralidad de agrupamientos se basa en sellos de tiempo de unidades respectivas de datos, indicando los sellos de tiempo momentos respectivos en los que se actualizaron por última vez las unidades de datos.

EEE 31: Un procedimiento para obtener datos modificados de una fuente de datos, comprendiendo el procedimiento: obtener, mediante uno o más procesadores de soporte físico de ordenador, información que indica una pluralidad de agrupamientos de datos de una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determinar, mediante los uno o más procesadores de soporte físico de ordenador, un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento

y un número histórico de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; acceder, mediante los uno o más procesadores de soporte físico de ordenador, unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; comparar, mediante los uno o más procesadores de soporte físico de ordenador, las unidades de datos incluidas en el primer agrupamiento con unidades de datos de una versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento de la fuente de datos han sido modificadas, en el que la versión local correspondiente del primer agrupamiento de unidades de datos es una versión comprimida del primer agrupamiento de unidades de datos; extraer, mediante los uno o más procesadores de soporte físico de ordenador, las unidades modificadas de datos del primer agrupamiento; y remitir, mediante los uno o más procesadores de soporte físico de ordenador, las unidades modificadas extraídas de datos a un sistema de destino.

EEE 32: El procedimiento de la EEE 31, en el que la versión comprimida del primer agrupamiento de unidades de datos es una estructura probabilística de datos que ahorra espacio que incluye información acerca de las unidades de datos incluidas en el primer agrupamiento en un primer momento antes de dicha obtención de la información que indica la pluralidad de agrupamientos de los datos de la fuente de datos.

EEE 33: El procedimiento de la EEE 32, en el que la estructura probabilística de datos que ahorra espacio determina si una unidad particular de datos incluida en el primer agrupamiento fue incluida en el primer agrupamiento en un primer momento.

EEE 34: El procedimiento de la EEE 31 o de la EEE 32, en el que la estructura probabilística de datos que ahorra espacio es un filtro de Bloom.

EEE 35: El procedimiento de cualquiera de las EEE 31-34, en el que la versión comprimida de los datos no comprende una copia del primer agrupamiento.

EEE 36: El procedimiento de cualquiera de las EEE 31-35, en el que la versión local correspondiente de los datos se almacena en un primer almacenamiento y las unidades modificadas extraídas de datos remitidas al sistema de destino se almacenan en un segundo almacenamiento, teniendo el primer almacenamiento una menor capacidad de almacenamiento que el segundo almacenamiento.

EEE 37: El procedimiento de la EEE 36, en el que el primer almacenamiento es un almacenamiento conectado en red (NAS).

EEE 38: El procedimiento de cualquiera de las EEE 31-37, en el que la fuente de datos es una base de datos o un fichero.

EEE 39: El procedimiento de cualquiera de las EEE 31-38, en el que la pluralidad de agrupamientos se basa en sellos de tiempo de unidades respectivas de datos, indicando los sellos de tiempo los momentos respectivos en los que se actualizaron por última vez las unidades de datos.

EEE 40: Un medio transitorio o no transitorio legible por un ordenador que comprende instrucciones para obtener datos modificados de una fuente de datos que provoca que un procesador de ordenador: obtenga información que indica una pluralidad de agrupamientos de datos de una fuente de datos, indicando la información un número de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; determine un primer agrupamiento de la pluralidad de agrupamientos que incluye una o más unidades de datos que han sido modificadas comparando un primer número de unidades de datos incluidas en el primer agrupamiento y un número histórico de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos; acceda a unidades de datos incluidas en el primer agrupamiento procedente de la fuente de datos; compare las unidades de datos incluidas en el primer agrupamiento con unidades de datos de una versión local correspondiente del primer agrupamiento para determinar qué unidades de datos del primer agrupamiento procedente de la fuente de datos han sido modificadas, en el que la versión local correspondiente del primer agrupamiento de unidades de datos es una versión comprimida del primer agrupamiento de unidades de datos; extraiga las unidades modificadas de datos del primer agrupamiento; y remita las unidades modificadas extraídas de datos a un sistema de destino.

Se debería hacer hincapié en que se pueden realizar muchas variaciones y modificaciones a las realizaciones descritas anteriormente, debiendo entenderse que sus elementos se encuentran entre otros ejemplos aceptables. Se concibe que todas las modificaciones y variaciones tales estén incluidas en la presente memoria dentro del alcance de la presente divulgación. La anterior descripción detalla ciertas realizaciones de la invención. Sin embargo, se apreciará que no importa cuán detallado parezca lo anterior en el texto, la invención puede ser puesta en práctica de muchas formas. Como también se ha indicado anteriormente, se debería hacer notar que el uso de terminología particular cuando se describen ciertas características o aspectos de la invención no debería interpretarse que implique que la terminología está siendo redefinida en la presente memoria para estar restringida a incluir cualquier característica específica de los rasgos o aspectos de la invención con la que está asociada esa terminología. Por lo

tanto, se debería interpretar el alcance de la invención según las reivindicaciones adjuntas y cualquier equivalente de las mismas.

REIVINDICACIONES

1. Un procedimiento (400a, 400b) para obtener datos modificados (340) procedentes de una fuente (110, 210) de datos y remitir los datos modificados obtenidos a un sistema (170, 270) de destino, comprendiendo el procedimiento (400a):

obtener (401a, 401b), mediante uno o más procesadores (504) de soporte físico de ordenador, información (310) que indica una pluralidad de agrupamientos de datos almacenados en uno o más ficheros (110b, 210b) o bases (110a, 210a) de datos en una fuente (110, 210) de datos, en el que la pluralidad de agrupamientos se basa en la fecha y la hora de los sellos de tiempo de unidades respectivas de datos, de forma que las unidades de datos se distribuyan de manera no uniforme en la pluralidad de agrupamientos, cada uno de los agrupamientos asociado con un distinto atributo de unidad de datos, de forma que cada uno de los agrupamientos respectivos incluya unidades de datos que tienen el atributo correspondiente de unidad de datos, en el que la información (310) indica una cantidad actual de unidades de datos incluidas en cada uno de la pluralidad de agrupamientos en un momento actual, en el que se obtiene la información (310) utilizando una consulta de un atributo de sello de tiempo de la unidad de datos;

acceder, mediante los uno o más procesadores (504) de soporte físico de ordenador, información histórica (315) de agrupamiento que indica una cantidad histórica de unidades de datos incluidas en al menos algunos de la pluralidad de agrupamientos en un momento anterior, en el que para un primer subconjunto (320) de los agrupamientos una cantidad actual de unidades de datos en el momento actual y una cantidad histórica de unidades de datos no son idénticas, y para otro subconjunto de los agrupamientos una cantidad actual de unidades de datos y una cantidad histórica de unidades de datos son idénticas;

comparar, mediante los uno o más procesadores (504) de soporte físico de ordenador, la cantidad de unidades de datos en cada uno de los agrupamientos correspondientes en el momento anterior y el momento actual;

determinar (402a, 402b), mediante los uno o más procesadores (504) de soporte físico de ordenador para cada agrupamiento (330) en el primer subconjunto (320), en función de dicha comparación, que una cantidad actual de unidades de datos incluidas en un agrupamiento correspondiente (330) y una cantidad histórica de unidades de datos incluidas en un agrupamiento histórico (335) de unidades de datos no son idénticas, lo que indica que una o más unidades de datos del agrupamiento correspondiente (330) han sido modificadas; y

en respuesta a la determinación de que la cantidad actual de unidades de datos es mayor o menor que la cantidad histórica de unidades de datos, se llevan a cabo las siguientes etapas del procedimiento para cada agrupamiento (330) en el primer subconjunto (320);

acceder (403a, 403b), mediante los uno o más procesadores (504) de soporte físico de ordenador, a unidades de datos incluidas en un agrupamiento (330) en el primer subconjunto (320) procedente de la fuente (110, 210) de datos;

comparar (404a, 404b), mediante los uno o más procesadores (504) de soporte físico de ordenador, las unidades de datos incluidas en el agrupamiento (330) en el primer subconjunto (320) con unidades de datos de una versión local correspondiente del agrupamiento (335) para determinar qué unidades de datos del agrupamiento (330) en el primer subconjunto (320) procedente de la fuente (110, 210) de datos han sido modificadas;

extraer (405a, 405b), mediante los uno o más procesadores (504) de soporte físico de ordenador, las unidades modificadas (340) de datos del agrupamiento (330) en el primer subconjunto (320); y

remitir, mediante los uno o más procesadores (504) de soporte físico de ordenador, las unidades modificadas extraídas (340) de datos al sistema (170, 270) de destino.

2. El procedimiento (400a, 400b) de la Reivindicación 1,

en el que la pluralidad de agrupamientos se basa en los sellos de tiempo de las unidades respectivas de datos, en el que los sellos de tiempo indican momentos respectivos en los que se actualizaron por última vez las unidades de datos.

3. El procedimiento (400a, 400b) de la Reivindicación 2,

en el que los sellos de tiempo de las unidades respectivas de datos incluyen una fecha y un momento en el que se actualizaron por última vez las unidades respectivas de datos, y en el que la pluralidad de agrupamientos se basa únicamente en la fecha de los sellos de tiempo de las unidades respectivas de datos.

4. El procedimiento (400a, 400b) de la Reivindicación 3,

en el que cada uno de la pluralidad de agrupamientos está asociado con una fecha distinta y/o las fechas asociadas con cada uno de la pluralidad de agrupamientos se encuentran en un intervalo de fechas.

5. El procedimiento (400a, 400b) de cualquiera de las Reivindicaciones 1-4,

en el que, en respuesta a la determinación de que la cantidad actual de unidades de datos es superior a la cantidad histórica correspondiente de unidades de datos, el procedimiento comprende, además:

identificar las unidades modificadas (340) de datos como unidades añadidas o actualizadas de datos; y
remitir las unidades modificadas (340) de datos al sistema (170, 270) de destino para ser almacenadas,

y/o

en el que, en respuesta a la determinación de que la cantidad actual de unidades de datos es menor que la cantidad histórica correspondiente de unidades de datos, el procedimiento comprende, además:

identificar las unidades modificadas (340) de datos como unidades borradas de datos; y
remitir las unidades modificadas (340) de datos al sistema (170, 270) de destino para ser eliminadas.

6. El procedimiento (400a, 400b) de cualquiera de las Reivindicaciones 1-5,

en el que el procedimiento comprende, además:

asignar un identificador único a cada una de las unidades de datos incluidas en cada agrupamiento (330) en el primer subconjunto, y

determinar si una primera unidad modificada de datos de las unidades modificadas (340) de datos es una nueva unidad de datos o una unidad actualizada de datos en función del identificador único asociado con la primera unidad modificada de datos,

y/o

en el que una unidad de datos incluida en un agrupamiento (330) en el primer subconjunto es una fila en las una o más bases (110a, 210a) de datos o una línea en los uno o más ficheros (110b, 210b).

7. El procedimiento (400a, 400b) de cualquiera de las Reivindicaciones 1-6,

en el que el procedimiento comprende, además:

obtener (401a, 401b) la información (310) que indica la pluralidad de agrupamientos de los datos almacenados en los uno o más ficheros (110b, 210b) en la fuente (110, 210) de datos utilizando un primer adaptador, y

obtener (401a, 401b) la información (310) que indica la pluralidad de agrupamientos de los datos almacenados en las una o más bases (110a, 210a) de datos en la fuente (110, 210) de datos utilizando un segundo adaptador,

y/u

obtener (401a, 401b) la información (310) que indica la pluralidad de agrupamientos de los datos en un intervalo,

y/u

obtener (401a, 401b) la información (310) que indica la pluralidad de agrupamientos de los datos en respuesta a la recepción de una solicitud procedente del sistema (170, 270) de destino.

8. El procedimiento (400a, 400b) de cualquiera de las Reivindicaciones 1-7,

en el que la versión local correspondiente del agrupamiento (335) comprende una copia de unidades de datos incluidas en el agrupamiento (335) en un primer momento antes de dicha obtención (401a) de la información (310) que indica la pluralidad de agrupamientos de los datos,

y/o

en el que una unidad de datos incluida en el agrupamiento (330) en el primer subconjunto (320) es una fila en las una o más bases (110a, 210a) de datos o una línea en los uno o más ficheros (110b, 210b).

9. El procedimiento (400b) de cualquiera de las Reivindicaciones 1-8, en el que la versión local correspondiente del agrupamiento (330) de unidades de datos es una versión comprimida del agrupamiento (330) de unidades de datos.

10. El procedimiento (400b) de la Reivindicación 9, en el que la versión comprimida del agrupamiento (330) de unidades de datos es una estructura probabilística de datos que ahorra espacio que incluye información (315) acerca de unidades de datos incluidas en el agrupamiento (335) en un primer momento antes de dicha obtención (401b) de la información (310) que indica la pluralidad de agrupamientos de los datos de la fuente (110, 210) de datos.

11. El procedimiento (400b) de la Reivindicación 10,

en el que la estructura probabilística de datos que ahorra espacio determina si una unidad particular de datoss incluida en el agrupamiento (330) fue incluida en el agrupamiento (335) en el primer momento,

y/o

en el que la estructura probabilística de datos que ahorra espacio es un filtro de Bloom (255), en el que, en algunas realizaciones, se selecciona el filtro de Bloom (255) entre una pluralidad de filtros de Bloom, cada uno de los cuales puede incluir un número distinto de unidades de datos.

12. El procedimiento (400b) de cualquiera de las Reivindicaciones 9-11,

en el que la versión comprimida de los datos no comprende una copia del agrupamiento (330),

y/o

en el que la versión local correspondiente de los datos se almacena en un primer almacenamiento y las unidades modificadas extraídas (340) de datos remitidas al sistema (170, 270) de destino son almacenadas en un segundo almacenamiento, teniendo el primer almacenamiento una menor capacidad de almacenamiento que el segundo almacenamiento, en el que, en algunas realizaciones, el primer almacenamiento es un almacenamiento conectado en red (NAS).

13. El procedimiento (400b) de cualquiera de las Reivindicaciones 9-12,

en el que la fuente (110, 210) de datos es una base (110a, 210a) de datos o un fichero (110b, 210b),

y/o

en el que la pluralidad de agrupamientos se basa en sellos de tiempo de unidades respectivas de datos, indicando los sellos de tiempo momentos respectivos en los que se actualizaron por última vez las unidades de datos.

14. Un sistema (100, 200a, 200b) de ordenador configurado para obtener datos modificados (340) procedentes de una fuente (110, 210) de datos y remitir los datos modificados obtenidos a un sistema (170, 270) de destino, comprendiendo el sistema (100, 200a, 200b) de ordenador:

uno o más procesadores (504) de soporte físico de ordenador configurados para ejecutar código para provocar que el sistema (100, 200a, 200b) lleve a cabo operaciones que incluyen las operaciones enumeradas en cualquiera de las Reivindicaciones 1-13.

15. Un medio legible por un ordenador que comprende instrucciones para obtener datos modificados (340) procedentes de una fuente (110, 210) de datos que provocan que un procesador (504) de ordenador lleve a cabo operaciones que incluyen las operaciones enumeradas en cualquiera de las Reivindicaciones 1-13.

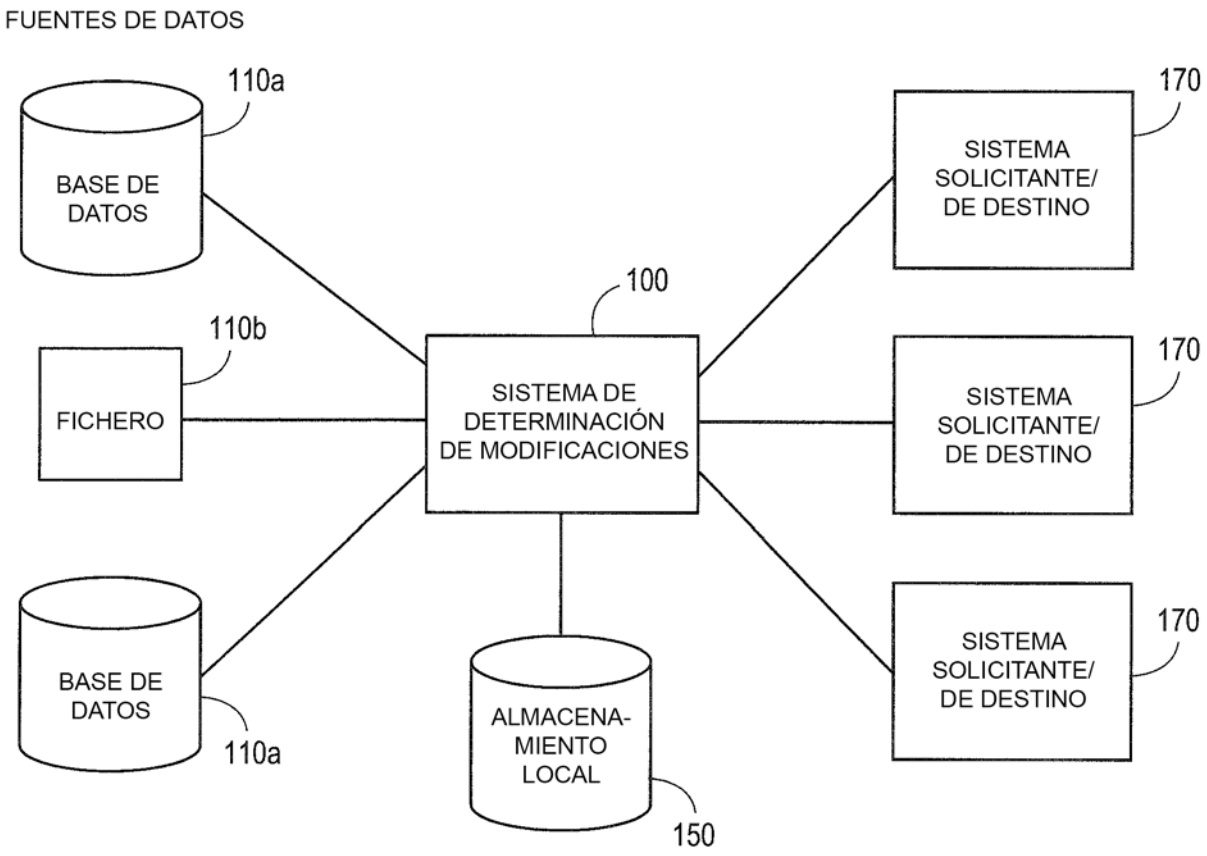
FUENTES DE DATOS
110a
BASE DE DATOS
110b
FICHERO
BASE DE DATOS
110a
100
SISTEMA DE DETERMINACIÓN DE MODIFICACIONES
ALMACENA-MIENTO LOCAL
150
170
SISTEMA SOLICITANTE/ DE DESTINO
170
SISTEMA SOLICITANTE/ DE DESTINO
170
SISTEMA SOLICITANTE/ DE DESTINO

FIG. 1

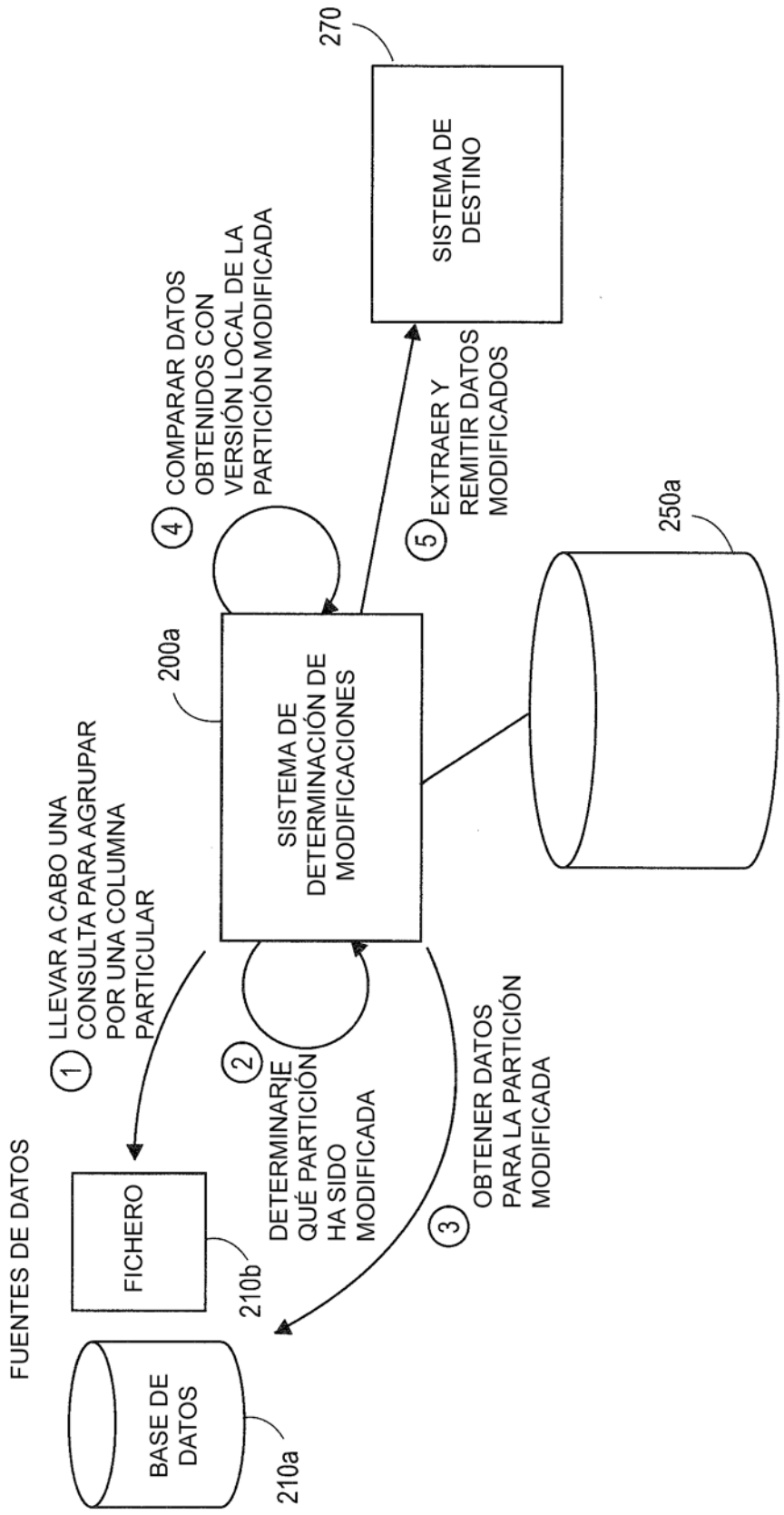
FUENTES DE DATOS
BASE DE DATOS
210a
FICHERO
210b
1 LLEVAR A CABO UNA CONSULTA PARA AGRUPAR POR UNA COLUMNA PARTICULAR
2 DETERMINARIE QUÉ PARTICIÓN HA SIDO MODIFICADA
3 OBTENER DATOS PARA LA PARTICIÓN MODIFICADA
SISTEMA DE DETERMINACIÓN DE MODIFICACIONES
200a
4 COMPARAR DATOS OBTENIDOS CON VERSIÓN LOCAL DE LA PARTICIÓN MODIFICADA
5 EXTRAER Y REMITIR DATOS MODIFICADOS
SISTEMA DE DESTINO
270
250a

FIG. 2A

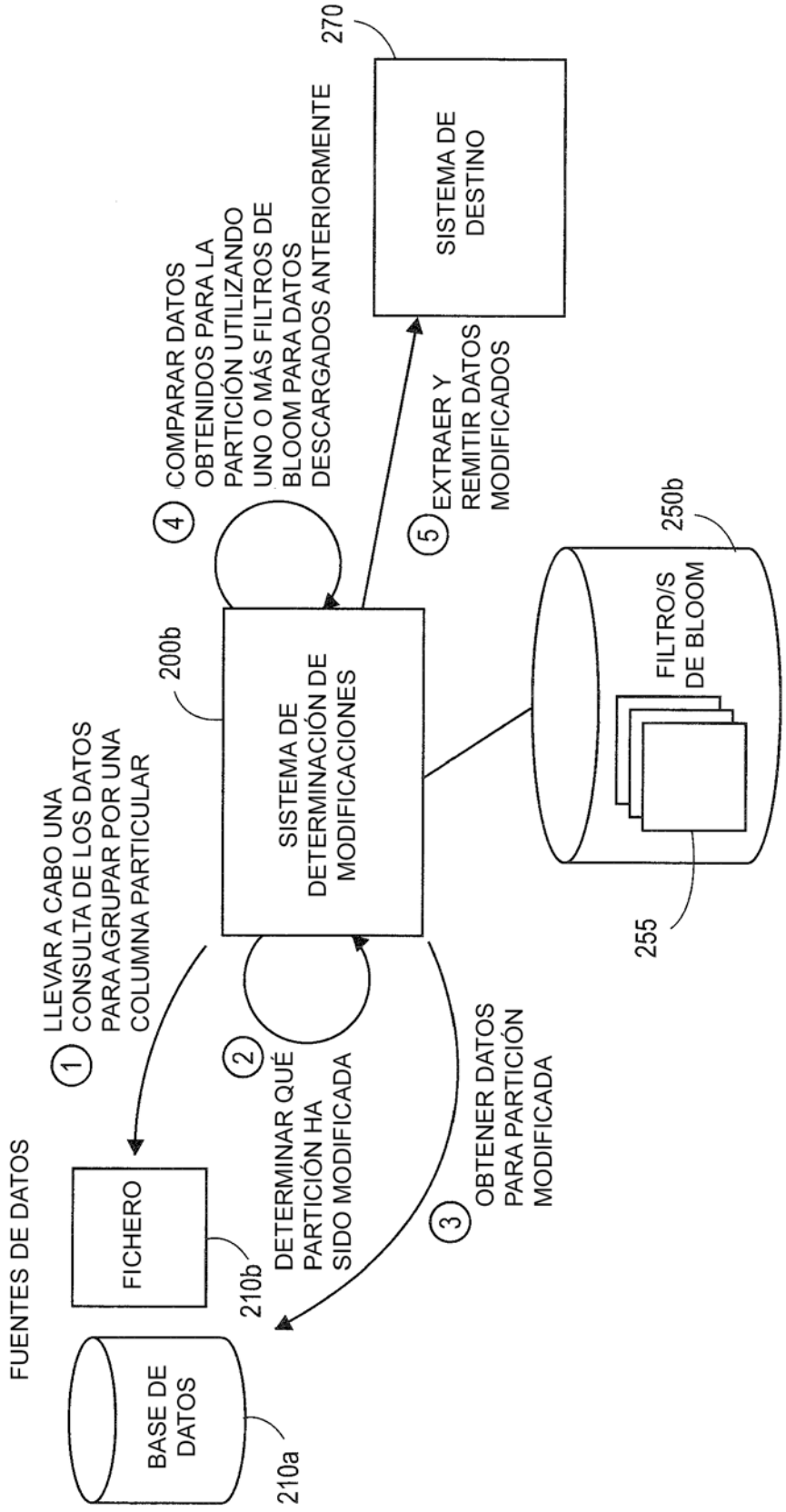
FUENTES DE DATOS
BASE DE DATOS
210a
FICHERO
210b
SISTEMA DE DETERMINACIÓN DE MODIFICACIONES
200b
SISTEMA DE DESTINO
270
FILTRO/S DE BLOOM
250b
255
1 LLEVAR A CABO UNA CONSULTA DE LOS DATOS PARA AGRUPAR POR UNA COLUMNA PARTICULAR
2 DETERMINAR QUÉ PARTICIÓN HA SIDO MODIFICADA
3 OBTENER DATOS PARA PARTICIÓN MODIFICADA
4 COMPARAR DATOS OBTENIDOS PARA LA PARTICIÓN UTILIZANDO UNO O MÁS FILTROS DE BLOOM PARA DATOS DESCARGADOS ANTERIORMENTE
5 EXTRAER Y REMITIR DATOS MODIFICADOS

FIG. 2B

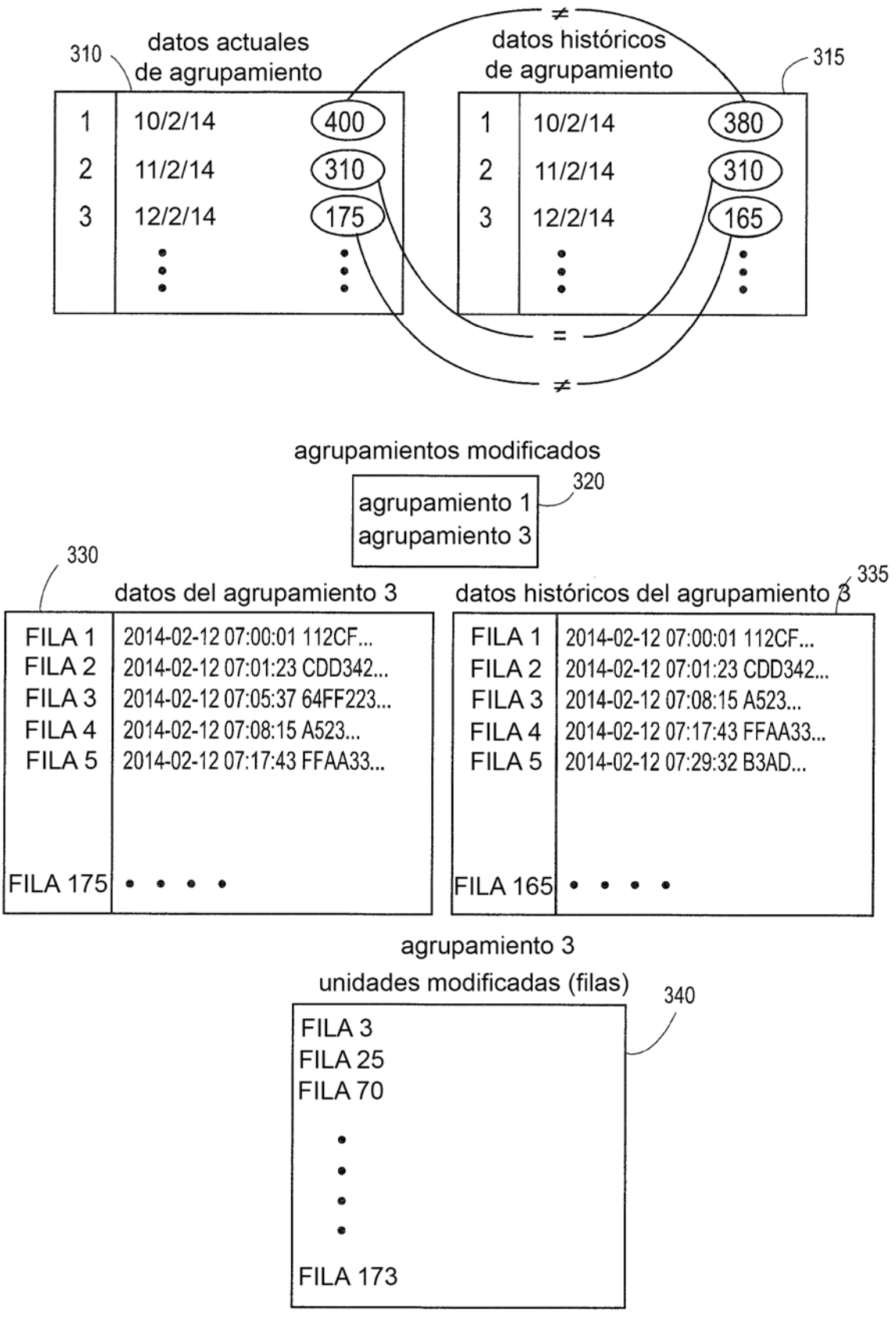
310
datos actuales
de agrupamiento
1 10/2/14 400
2 11/2/14 310
3 12/2/14 175
≠
datos históricos
de agrupamiento
315
1 10/2/14 380
2 11/2/14 310
3 12/2/14 165
=
≠
agrupamientos modificados
320
agrupamiento 1
agrupamiento 3
330
datos del agrupamiento 3
FILA 1 2014-02-12 07:00:01 112CF...
FILA 2 2014-02-12 07:01:23 CDD342...
FILA 3 2014-02-12 07:05:37 64FF223...
FILA 4 2014-02-12 07:08:15 A523...
FILA 5 2014-02-12 07:17:43 FFAA33...
FILA 175
335
datos históricos del agrupamiento 3
FILA 1 2014-02-12 07:00:01 112CF...
FILA 2 2014-02-12 07:01:23 CDD342...
FILA 3 2014-02-12 07:08:15 A523...
FILA 4 2014-02-12 07:17:43 FFAA33...
FILA 5 2014-02-12 07:29:32 B3AD...
FILA 165
agrupamiento 3
unidades modificadas (filas)
340
FILA 3
FILA 25
FILA 70
FILA 173

FIG. 3

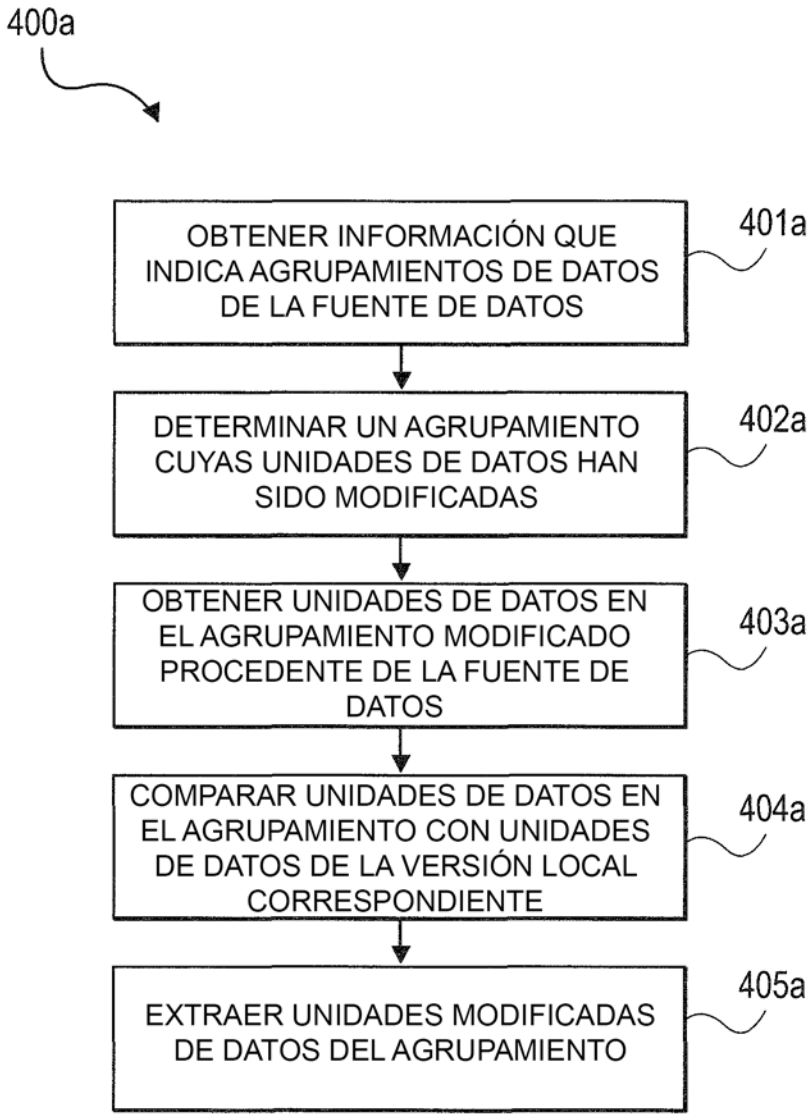
400a
OBTENER INFORMACIÓN QUE INDICA AGRUPAMIENTOS DE DATOS DE LA FUENTE DE DATOS
401a
DETERMINAR UN AGRUPAMIENTO CUYAS UNIDADES DE DATOS HAN SIDO MODIFICADAS
402a
OBTENER UNIDADES DE DATOS EN EL AGRUPAMIENTO MODIFICADO PROCEDENTE DE LA FUENTE DE DATOS
403a
COMPARAR UNIDADES DE DATOS EN EL AGRUPAMIENTO CON UNIDADES DE DATOS DE LA VERSIÓN LOCAL CORRESPONDIENTE
404a
EXTRAER UNIDADES MODIFICADAS DE DATOS DEL AGRUPAMIENTO
405a

FIG. 4A

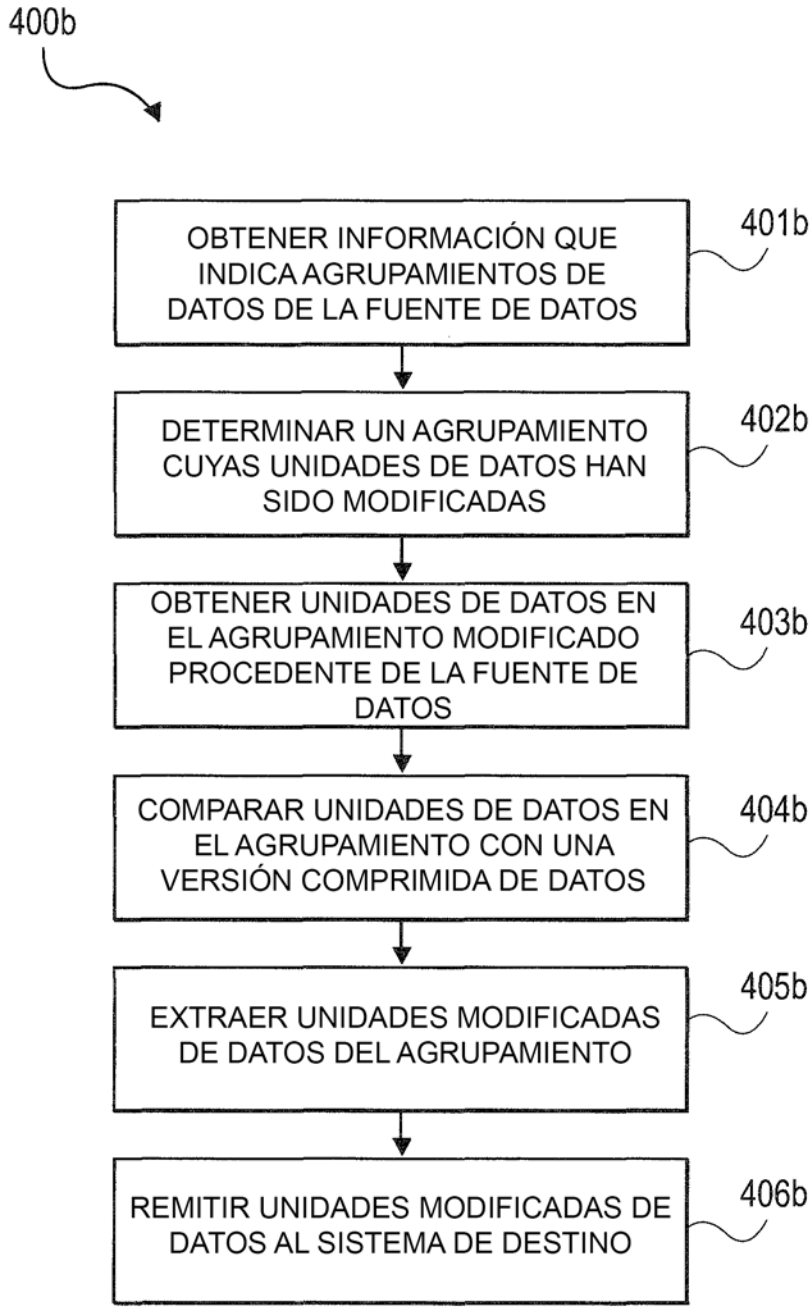
400b
OBTENER INFORMACIÓN QUE INDICA AGRUPAMIENTOS DE DATOS DE LA FUENTE DE DATOS
401b
DETERMINAR UN AGRUPAMIENTO CUYAS UNIDADES DE DATOS HAN SIDO MODIFICADAS
402b
OBTENER UNIDADES DE DATOS EN EL AGRUPAMIENTO MODIFICADO PROCEDENTE DE LA FUENTE DE DATOS
403b
COMPARAR UNIDADES DE DATOS EN EL AGRUPAMIENTO CON UNA VERSIÓN COMPRIMIDA DE DATOS
404b
EXTRAER UNIDADES MODIFICADAS DE DATOS DEL AGRUPAMIENTO
405b
REMITIR UNIDADES MODIFICADAS DE DATOS AL SISTEMA DE DESTINO
406b

FIG. 4B

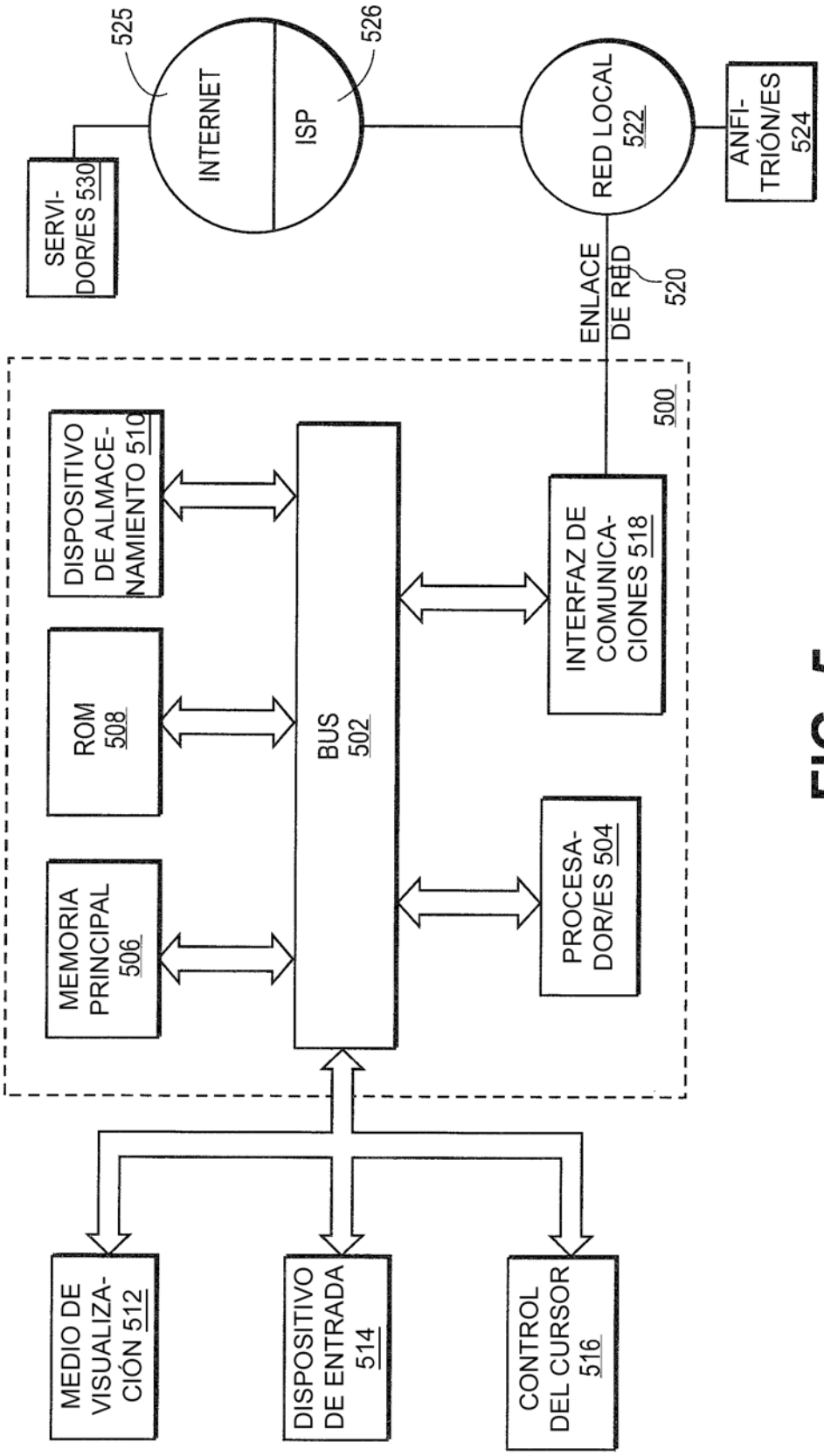
525
INTERNET
ISP
526
SERVI-
DOR/ES 530
RED LOCAL
522
ANFI-
TRIÓN/ES
524
ENLACE
DE RED
520
500
DISPOSITIVO
DE ALMACE-
NAMIENTO 510
ROM
508
MEMORIA
PRINCIPAL
506
BUS
502
INTERFAZ DE
COMUNICA-
CIONES 518
PROCESA-
DOR/ES 504
MEDIO DE
VISUALIZA-
CIÓN 512
DISPOSITIVO
DE ENTRADA
514
CONTROL
DEL CURSOR
516

FIG. 5