(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 648 972**

(51) Int. CI.:

G06F 11/10 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **07.08.2013 PCT/CN2013/080990**

(87) Fecha y número de publicación internacional: **14.08.2014 WO14121593**

(96) Fecha de presentación y número de la solicitud europea: **07.08.2013 E 13801457 (6)**

(97) Fecha y número de publicación de la concesión europea: **11.10.2017 EP 2787446**

(54) Título: **Procedimiento, dispositivo y sistema de almacenamiento distribuido**

(30) Prioridad:

08.02.2013 CN 201310050257

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
09.01.2018

(73) Titular/es:

HUAWEI TECHNOLOGIES CO., LTD. (100.0%)
Huawei Administration Building, Bantian
Longgang District
Shenzhen, Guangdong 518129, CN

(72) Inventor/es:

FENG, BIN;
HUANG, CHENG y
GONG, XUEWEN

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Procedimiento, dispositivo y sistema de almacenamiento distribuido

SECTOR TÉCNICO

La presente invención se refiere al sector técnico del almacenamiento de datos, y en particular, a un procedimiento, un aparato y un sistema de almacenamiento distribuido.

ANTECEDENTES

En un sistema de almacenamiento distribuido, con el objetivo de conseguir fiabilidad, se adopta una tecnología de redundancia de múltiples copias a nivel de archivo, o una tecnología de codificación con redundancia a nivel de bloque de datos, por ejemplo, una tecnología de codificación de borrado (codificación de borrado). En la tecnología de redundancia de múltiples copias, la probabilidad de pérdida de datos se puede reducir almacenando múltiples copias iguales de un archivo de datos, y en la tecnología de codificación con redundancia, la fiabilidad se puede mejorar añadiendo un bloque de verificación para cualesquiera datos parciales en un archivo.

En general, se puede adoptar una tabla de hash distribuida (distributed hash table, DHT) para almacenar un bloque de datos y un bloque de verificación. Sin embargo, debido a la aleatoriedad de la DHT, no se puede evitar que múltiples bloques de datos de un mismo segmento de datos se desplieguen en un mismo nodo de almacenamiento físico, y por lo tanto no se puede evitar que la invalidez de un solo nodo de almacenamiento físico (por ejemplo, un bastidor, un servidor o un disco duro) tenga como resultado un riesgo de pérdida de datos. Por ejemplo, se adopta una tecnología de codificación de borrado M+N, donde M es el número de bloques de datos y N es el número de bloques de verificación, y cuándo se despliegan más de N+1 bloques de datos o bloques de verificación en un mismo disco duro, un fallo del disco duro puede tener como resultado una pérdida de los M bloques de datos, y por lo tanto puede tener como resultado una no disponibilidad de todo el archivo. Utilizando como ejemplo un mecanismo de almacenamiento con redundancia 12 + 3, cuando se pierden más de 4 bloques de datos, un segmento de datos se puede perder y no puede ser restablecido.

En otras palabras, en un sistema de almacenamiento distribuido existente, un fallo en un único punto (por ejemplo, un disco duro, un servidor o un bastidor) puede tener como resultado una pérdida de datos, y el riesgo y la probabilidad del fallo son extremadamente elevados, especialmente cuando la escala del sistema de almacenamiento distribuido es relativamente pequeña, reduciendo por lo tanto la fiabilidad del sistema de almacenamiento distribuido.

El documento WO 00/27108 A1 da a conocer un sistema de almacenamiento, tal como un sistema de video a la carta, que comprende una serie de unidades de almacenamiento, tales como unidades de disco. Cada una de las unidades de almacenamiento tiene una serie de zonas de almacenamiento con tiempos previstos de recuperación de datos diferentes entre sí. Los datos se almacenan en unidades de datos, comprendiendo cada unidad de datos N bloques, donde N ≥2. Los bloques de la unidad de datos están distribuidos sobre las unidades de almacenamiento. Los bloques comprenden información redundante, de tal modo que cualquiera de una serie de selecciones que comprenden N - K de los N bloques es suficiente para recuperar la unidad de datos, donde K≥1. Para recuperar una unidad de datos a partir de las unidades de almacenamiento, un lector determina qué N - K de los N bloques recuperar, en base a un procedimiento de selección.

RESUMEN

Las realizaciones de la presente invención dan a conocer un sistema de almacenamiento distribuido y un procedimiento, un aparato y un sistema de almacenamiento distribuido, que pueden mejorar la fiabilidad de un sistema de almacenamiento distribuido.

En un primer aspecto, se da a conocer un procedimiento de almacenamiento distribuido, que incluye: dividir un archivo de datos para generar K segmentos de datos, dividir cada segmento de datos de los K segmentos de datos para generar M bloques de datos para cada segmento de datos, y llevar a cabo una codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación; determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida; y almacenar por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en dichos por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros positivos.

Haciendo referencia al primer aspecto, en un primer posible modo de implementación, la determinación, utilizando un algoritmo aleatorio, de un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación incluye: llevar a cabo un cálculo de hash sobre un identificador de un bloque de datos o de un bloque de verificación de los M bloques de datos o de los N bloques de verificación para generar un valor de clave; y determinar, de acuerdo con el valor de clave, un nodo

de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utilizar el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

Haciendo referencia al primer aspecto, en un segundo posible modo de implementación, la determinación, utilizando un algoritmo aleatorio, de un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación incluye: llevar a cabo un cálculo de hash sobre un identificador del segmento de datos que está dividido en los M bloques de datos, para generar un valor de clave; y determinar, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utilizar el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

Haciendo referencia al primer aspecto o a cualquiera de los posibles modos de implementación anteriores, en un tercer posible modo de implementación, la determinación de por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento incluye: determinar M+N diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento; y almacenar por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos incluye: almacenar los M bloques de datos y los N bloques de verificación en los M+N diferentes nodos de almacenamiento físicos.

Haciendo referencia al primer aspecto o a cualquiera de los posibles modos de implementación anteriores, en un cuarto posible modo de implementación, la determinación de por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, incluye: mapear el primer nodo de almacenamiento físico a un bloque de los por lo menos M+1 bloques, y determinar, en base a una posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a otros por lo menos M bloques de los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos incluyen el primer nodo de almacenamiento físico; o determinar, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos no incluyen el primer nodo de almacenamiento físico.

Haciendo referencia al primer aspecto o a cualquiera de los posibles modos de implementación anteriores, en un quinto posible modo de implementación, los diferentes nodos de almacenamiento físicos son discos duros, y el número de discos duros es mayor o igual que M+1; o los diferentes nodos de almacenamiento físicos son servidores, donde el número de servidores es mayor o igual que M+1; o los diferentes nodos de almacenamiento físicos son bastidores, donde el número de bastidores es mayor o igual que M+1.

Haciendo referencia al primer aspecto o a cualquiera de los posibles modos de implementación anteriores, en un sexto posible modo de implementación, el procedimiento del primer aspecto incluye además: en caso de que uno de los por lo menos M+1 diferentes nodos de almacenamiento físicos que almacenan por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación falle, restablecer los datos que se encuentran en el nodo de almacenamiento físico que falla de los por lo menos M+1 diferentes nodos de almacenamiento físicos en un nodo de almacenamiento de respaldo en caliente en un sistema de almacenamiento distribuido en el que está situado el nodo de almacenamiento físico.

Haciendo referencia al primer aspecto o a cualquiera de los posibles modos de implementación anteriores, en un séptimo posible modo de implementación, el procedimiento del primer aspecto incluye además: en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle y si L no es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es opuesto al sentido de una secuencia en el primer modo de ordenamiento; y si L es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques posteriores al bloque L-ésimo, a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es el mismo que el sentido de la secuencia en el primer modo de ordenamiento; o en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido; o en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el

primer modo de ordenamiento y los bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Haciendo referencia al primer aspecto o a cualquiera del primer al sexto posibles modos de implementación, en un octavo posible modo de implementación, el procedimiento del primer aspecto incluye además: en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos y si P no es mayor que (M+N)/2, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es el mismo que el sentido de una secuencia en el primer modo de ordenamiento; y si P es mayor que (M+N)/2, migrar un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es opuesto al sentido de la secuencia en el primer modo de ordenamiento, y P es un entero; o en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido; o en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico en el que está situado un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y los N

bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Haciendo referencia al primer aspecto o a cualquiera de los posibles modos de implementación anteriores, en un noveno posible modo de implementación, el procedimiento del primer aspecto incluye además: cuando se tiene que leer el archivo de datos, determinar, utilizando el algoritmo aleatorio, el primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento; leer por lo menos M bloques desde los por lo menos M diferentes nodos de almacenamiento físicos, donde los M bloques leídos incluyen los M bloques de datos o incluyen algunos bloques de datos de los M bloques de datos y algunos bloques de verificación de los N bloques de verificación; y llevar a cabo una descodificación y una verificación inversa sobre los por lo menos M bloques para generar M bloques de datos descodificados, y combinar los M bloques de datos descodificados para generar el archivo de datos.

Haciendo referencia al primer aspecto o a cualquiera de los anteriores posibles modos de implementación, en un décimo posible modo de implementación, el archivo de datos se obtiene desde fuera del sistema de almacenamiento distribuido, o el archivo de datos se obtiene desde algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido.

Haciendo referencia al primer aspecto o a cualquiera de los anteriores posibles modos de implementación, en un undécimo posible modo de implementación, el algoritmo de redundancia es un algoritmo de codificación de borrado, y el algoritmo aleatorio es un algoritmo de tabla de hash distribuida.

Haciendo referencia al primer aspecto o a cualquiera de los anteriores posibles modos de implementación, en un duodécimo posible modo de implementación, los nodos de almacenamiento físicos diferentes son nodos de almacenamiento físicos diferentes en el sistema de almacenamiento distribuido, cada nodo de almacenamiento físico de los nodos de almacenamiento físicos diferentes incluye múltiples nodos de almacenamiento virtuales, y se despliegan nodos de almacenamiento virtuales con números de serie consecutivos en los diferentes nodos de almacenamiento físicos de acuerdo con un segundo modo de ordenamiento que es conforme con una regla preestablecida; la determinación, utilizando un algoritmo aleatorio, de un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y la determinación de por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento, incluye: determinar,

utilizando el algoritmo aleatorio, el número de serie de un primer nodo de almacenamiento virtual correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación, y determinar números de serie de por lo menos M+1 nodos de almacenamiento virtuales en base al número de serie del primer nodo de almacenamiento virtual y de acuerdo con el primer modo de ordenamiento; y almacenar por separado por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos incluye: almacenar los por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en nodos de almacenamiento virtuales correspondientes a los números de serie de los por lo menos M+1 nodos de almacenamiento virtuales.

Haciendo referencia al primer aspecto o a cualquiera de los posibles modos de implementación anteriores, en un decimotercer posible modo de implementación, el primer modo de ordenamiento o el segundo modo de ordenamiento se refieren a un procedimiento de ordenamiento secuencial a un intervalo fijo.

En un segundo aspecto, se da a conocer un aparato de almacenamiento distribuido, que incluye: un módulo de generación, configurado para dividir un archivo de datos para generar K segmentos de datos, dividir cada segmento de datos de los K segmentos de datos para generar M bloques de datos para cada segmento de datos, y llevar a cabo codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación; un módulo de determinación, configurado para determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida; y un módulo de almacenamiento, configurado para almacenar por separado por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros positivos.

Haciendo referencia al segundo aspecto, en un primer posible modo de implementación del segundo aspecto, el módulo de determinación lleva a cabo un cálculo de hash sobre un identificador de un bloque de datos o de un bloque de verificación de los M bloques de datos o de los N bloques de verificación para generar un valor de clave, determina, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utiliza el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

Haciendo referencia al segundo aspecto, en un segundo posible modo de implementación del segundo aspecto, el módulo de determinación está configurado para llevar a cabo un cálculo de hash sobre un identificador del segmento de datos que está dividido en los M bloques de datos, con el fin de generar un valor de clave; determinar, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utilizar el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un tercer posible modo de implementación, el módulo de determinación determina M+N diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, y el módulo de almacenamiento almacena los M bloques de datos y los N bloques de verificación en los M+N diferentes nodos de almacenamiento físicos.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un cuarto posible modo de implementación, el módulo de determinación mapea el primer nodo de almacenamiento físico a un bloque de los por lo menos M+1 bloques, y determina, en base a una posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a los otros por lo menos M bloques de los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos incluyen el primer nodo de almacenamiento físico; o determina, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos no incluyen el primer nodo de almacenamiento físico.

Haciendo referencia al segundo aspecto o a cualquiera de los posibles modos de implementación anteriores del segundo aspecto, en un quinto posible modo de implementación, los diferentes nodos de almacenamiento físicos son discos duros, y el número de discos duros es mayor o igual que M+1; o los diferentes nodos de almacenamiento físicos son servidores, donde el número de servidores es mayor o igual que M+1; o los diferentes nodos de almacenamiento físicos son bastidores, donde el número de bastidores es mayor o igual que M+1.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un sexto posible modo de implementación, el aparato del segundo aspecto incluye además: un módulo de restablecimiento, configurado para, en caso de que uno de los por lo menos M+1 diferentes nodos de almacenamiento físicos que almacenan por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación falle, restablecer datos que se encuentran en el nodo de almacenamiento físico que falla de los por lo

menos M+1 diferentes nodos de almacenamiento físicos en un nodo de almacenamiento de respaldo en caliente en un sistema de almacenamiento distribuido en el que está situado el nodo de almacenamiento físico.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un séptimo posible modo de implementación, el aparato del segundo aspecto incluye además: un módulo de restablecimiento, configurado para, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle, y si L no es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es opuesto al sentido de una secuencia en el primer modo de ordenamiento; y si L es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es igual que el sentido de la secuencia en el primer modo de ordenamiento; o un módulo de restablecimiento, configurado para, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido; o un módulo de restablecimiento, configurado para, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un séptimo posible modo de implementación, el aparato del segundo aspecto incluye además: un módulo de expansión de la capacidad, configurado para, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, y si P no es mayor que (M+N)/2, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es igual que el sentido de una secuencia en el primer modo de ordenamiento, y si P es mayor que (M+N)/2, migrar un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es opuesto al sentido de la secuencia en el primer modo de ordenamiento, y P es un entero; o un módulo de expansión de la capacidad, configurado para, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido; o un módulo de expansión de la capacidad, configurado para, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico en el que está situado un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un noveno posible modo de implementación, el aparato del segundo aspecto incluye además: que el módulo de determinación está configurado además para, cuando se tiene que leer el archivo de datos, determinar, utilizando el algoritmo aleatorio, el primer nodo de almacenamiento físico correspondiente a un bloque

de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento; y el aparato del segundo aspecto incluye además: un módulo de lectura, configurado para leer por lo menos M bloques desde los M diferentes nodos de almacenamiento físicos, donde los M bloques leídos incluyen los M bloques de datos o incluyen algunos bloques de datos de los M bloques de datos y algunos bloques de verificación de los N bloques de verificación, donde el módulo de generación está configurado además para llevar a cabo descodificación y una verificación inversa sobre por lo menos M bloques para generar M bloques de datos descodificados, y combinar los M bloques de datos descodificados para generar el archivo de datos.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un décimo posible modo de implementación, el aparato del segundo aspecto incluye además: un módulo de adquisición, configurado para adquirir el archivo de datos desde el exterior del sistema de almacenamiento distribuido, o adquirir el archivo de datos desde algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un undécimo posible modo de implementación, el algoritmo de redundancia es un algoritmo de codificación de borrado, y el algoritmo aleatorio es un algoritmo de tabla de hash distribuida.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un duodécimo posible modo de implementación, los diferentes nodos de almacenamiento físicos son diferentes nodos de almacenamiento físicos en el sistema de almacenamiento distribuido, cada nodo de almacenamiento físico de los diferentes nodos de almacenamiento físicos incluye múltiples nodos de almacenamiento virtuales, y se despliegan nodos de almacenamiento virtuales con números de serie consecutivos en los diferentes nodos de almacenamiento físicos de acuerdo con un segundo modo de ordenamiento que es conforme con una regla preestablecida; el módulo de determinación determina, utilizando el algoritmo aleatorio, el número de serie de un primer nodo de almacenamiento virtual correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación, y determina números de serie de por lo menos M+1 nodos de almacenamiento virtuales en base al número de serie del primer nodo de almacenamiento virtual y de acuerdo con el primer modo de ordenamiento; y el módulo de almacenamiento almacena por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en nodos de almacenamiento virtuales correspondientes a los números de serie de los por lo menos M+1 nodos de almacenamiento virtuales.

Haciendo referencia al segundo aspecto o a cualquiera de los anteriores posibles modos de implementación del segundo aspecto, en un decimotercer posible modo de implementación, el primer modo de ordenamiento o el segundo modo de ordenamiento se refieren a un procedimiento de ordenamiento secuencial a un intervalo fijo.

En un tercer aspecto, se da a conocer un sistema de almacenamiento distribuido, que incluye: un cliente; múltiples nodos de almacenamiento físicos; y el aparato de almacenamiento distribuido del segundo aspecto, donde el aparato de almacenamiento distribuido lee/almacena un archivo de datos de un usuario en/desde los múltiples nodos de almacenamiento físicos, de acuerdo con una solicitud de almacenamiento/lectura enviada por el usuario a través del cliente.

En las realizaciones de la presente invención, se puede determinar un nodo de almacenamiento correspondiente a un bloque de datos de un archivo de datos utilizando un algoritmo aleatorio, se determinan por lo menos M+1 diferentes nodos de almacenamiento físicos en base al nodo de almacenamiento determinado y de acuerdo con un modo de ordenamiento basado en reglas, y por lo menos M+1 bloques de M bloques de datos y N bloques de verificación se almacenan en los por lo menos M+1 diferentes nodos de almacenamiento físicos, de tal modo que se pueden desplegar bloques de datos de segmentos de datos en nodos de almacenamiento físicos lo más diferentes posibles de acuerdo con el modo de ordenamiento basado en reglas, y se reduce la pérdida de datos que puede resultar de un fallo en un punto de cantar, mejorando de ese modo la fiabilidad de un sistema de almacenamiento distribuido.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Para describir más claramente las soluciones técnicas en las realizaciones de la presente invención, a continuación se describen brevemente los siguientes dibujos adjuntos necesarios para describir las realizaciones de la presente invención. Evidentemente, los dibujos adjuntos en la siguiente descripción muestran solamente algunas realizaciones de la presente invención, y un experto en la materia obtendrá sin esfuerzos creativos otros dibujos a partir de estos dibujos adjuntos.

La figura 1 es un diagrama de flujo esquemático de un procedimiento de almacenamiento distribuido, según una realización de la presente invención;

la figura 2 es un diagrama esquemático de un proceso en el que un sistema de almacenamiento distribuido divide y almacena un archivo de datos, según una realización de la presente invención;

la figura 3 es un diagrama de flujo esquemático de un proceso de almacenamiento distribuido, según una realización de la presente invención;

la figura 4 es un diagrama de flujo esquemático de un proceso de almacenamiento distribuido, según una realización de la presente invención;

la figura 5A es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos cuando un disco duro de un sistema de almacenamiento distribuido falla, según una realización de la presente invención;

la figura 5B es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos de un sistema de almacenamiento distribuido después de una recuperación de fallos, según una realización de la presente invención;

la figura 5C es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos de un sistema de almacenamiento distribuido, después de una recuperación de fallos, según otra realización de la presente invención;

la figura 6A es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos cuando se lleva a cabo expansión de la capacidad para un sistema de almacenamiento distribuido, según una realización de la presente invención;

la figura 6B es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos después de que se lleve a cabo expansión de la capacidad para un sistema de almacenamiento distribuido, según una realización de la presente invención;

la figura 7 es un diagrama estructural esquemático de un aparato de almacenamiento distribuido, según una realización de la presente invención;

la figura 8 es un diagrama estructural esquemático de un aparato de almacenamiento distribuido, según una realización de la presente invención;

la figura 9 es un diagrama estructural esquemático de un sistema de almacenamiento distribuido, según una realización de la presente invención;

la figura 10 es un diagrama de arquitectura de un sistema de almacenamiento distribuido, según una realización de la presente invención;

la figura 11A es un diagrama estructural esquemático de un servidor de almacenamiento/placa de almacenamiento de un sistema de almacenamiento distribuido, según una realización de la presente invención;

la figura 11B es un diagrama estructural esquemático de un disco duro de un sistema de almacenamiento distribuido, según una realización de la presente invención;

la figura 11C es un diagrama esquemático de un anillo lógico de nodos virtuales de un sistema de almacenamiento distribuido, según una realización de la presente invención; y

la figura 12 es un diagrama estructural esquemático de un aparato de almacenamiento distribuido, según otra realización de la presente invención.

DESCRIPCIÓN DE REALIZACIONES

A continuación se describen de manera clara y completa las soluciones técnicas de las realizaciones de la presente invención, haciendo referencia a los dibujos adjuntos en las realizaciones de la presente invención. Evidentemente, las realizaciones descritas son solamente una parte y no la totalidad de las realizaciones de la presente invención. Todas las demás realizaciones obtenidas sin esfuerzos creativos por un experto en la materia en base a las realizaciones de la presente invención caerán dentro del alcance de protección de la presente invención.

En un sistema de almacenamiento distribuido, se pueden montar múltiples discos duros en una matriz de almacenamiento, y el espacio de almacenamiento de cada disco duro se divide en múltiples nodos virtuales (nodo virtual, VN), que se denominan asimismo particiones (partición). Cada nodo virtual tiene un número de serie independiente, y los nodos virtuales tienen generalmente espacio de almacenamiento del mismo tamaño, por ejemplo, un nodo virtual con un número de serie lógico se despliega aleatoriamente en un disco duro con un número de serie físico.

Durante el almacenamiento de datos, el sistema de almacenamiento distribuido divide (dividir) un archivo para generar múltiples segmentos de datos (segmento de datos), a continuación divide un segmento de datos para generar múltiples bloques de datos (bloque de datos), añade bloques de verificación correspondientes, mapea los bloques de datos y los bloques de verificación a números de serie de nodo virtual utilizando una DHT, y finalmente distribuye aleatoriamente los bloques de datos y los bloques de verificación de datos en nodos virtuales correspondientes a los números de serie de nodo virtual. Análogamente, debido a la aleatoriedad de la DHT, sigue

siendo inevitable que múltiples bloques de datos de un mismo segmentos de datos se desplieguen en un mismo nodo de almacenamiento físico, y por lo tanto no se puede evitar que la no validez de un solo nodo de almacenamiento físico tenga como resultado un riesgo de pérdida de datos.

La figura 1 es un diagrama de flujo esquemático de un procedimiento de almacenamiento distribuido, según una realización de la presente invención. El procedimiento de la figura 1 puede ser ejecutado por un sistema de almacenamiento distribuido, y específicamente puede ser ejecutado por un motor de almacenamiento en el sistema de almacenamiento distribuido.

110. Dividir un archivo de datos para generar K segmentos de datos, dividir cada segmento de datos en los K segmentos de datos para generar M bloques de datos para cada segmento de datos y llevar a cabo codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación.

Por ejemplo, el sistema de almacenamiento distribuido puede recibir una solicitud de almacenamiento enviada por un usuario a través de un cliente, y autenticar al usuario según información de identificación del usuario, donde la información de identificación está contenida en la solicitud de almacenamiento. Esto no se limita a esta realización de la presente invención. Por ejemplo, el sistema de almacenamiento distribuido puede asimismo recibir directamente un archivo de datos cargado por el usuario y almacenarlo, sin necesidad de autenticación. Después de una autenticación satisfactoria, el sistema de almacenamiento distribuido puede recibir el archivo de datos enviado por el usuario.

Alternativamente, el archivo de datos se puede adquirir asimismo desde algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido, por ejemplo, cuando se lleva a cabo expansión de la capacidad a gran escala, un nuevo sistema de almacenamiento distribuido puede adquirir el archivo de datos desde algunos nodos de almacenamiento físicos (nodos de almacenamiento físicos de un sistema de almacenamiento distribuido original).

El sistema de almacenamiento distribuido de esta realización de la presente invención puede asignar un identificador para el archivo de datos recibido, y en caso de que la solicitud de almacenamiento incluya información del usuario y un tipo de servicio de almacenamiento, puede asignar asimismo un identificador para el archivo de datos de acuerdo con la información del usuario y el tipo de servicio de almacenamiento.

De acuerdo con esta realización de la presente invención, cuando el archivo de datos se divide en múltiples segmentos de datos, se puede añadir un identificador para cada segmento de datos. Cuando cada segmento de datos se divide en múltiples bloques de datos, se puede añadir un identificador para cada bloque de datos. Cuando la codificación de verificación se lleva a cabo sobre los bloques de datos utilizando un mecanismo de codificación con redundancia para generar un código de verificación, se puede añadir un identificador para cada bloque de verificación. Los tamaños de los segmentos de datos, los bloques de datos o los bloques de verificación pueden ser fijos o variables. Por ejemplo, el algoritmo de redundancia puede ser un algoritmo de codificación de borrado.

120. Determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida.

De acuerdo con esta realización de la presente invención, el algoritmo aleatorio puede ser un algoritmo de tabla de hash distribuida. Por ejemplo, puede estar preestablecida en una tabla de hash distribuida una relación de mapeo entre cada valor de clave y un número de serie de nodo de almacenamiento.

130. Almacenar por separado por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros.

De acuerdo con esta realización de la presente invención, M+1 bloques de datos y bloques de verificación se pueden desplegar en por lo menos M+1 nodos de almacenamiento, y otros bloques excepto los por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación se pueden desplegar aleatoriamente, de tal modo que cuando algún nodo de almacenamiento físico falla, se pueden restablecer los datos que se encuentran en el nodo de almacenamiento físico que falla. Esto no tiene como resultado una pérdida de datos, y mejora la fiabilidad del sistema de almacenamiento distribuido.

En esta realización de la presente invención, se puede determinar un nodo de almacenamiento físico correspondiente a un bloque de datos del archivo de datos utilizando el algoritmo aleatorio, se determinan por lo menos M+1 diferentes nodos de almacenamiento físicos en base al nodo de almacenamiento físico determinado y de acuerdo con un modo de ordenamiento basado en reglas, y se almacenan por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, de tal modo que los bloques de datos de los segmentos de datos se pueden desplegar en nodos de almacenamiento lo más diferentes posible de acuerdo con el modo de ordenamiento basado en reglas, y se reduce

la pérdida de datos que puede resultar de un fallo de un punto de cantar, mejorando de ese modo la fiabilidad del sistema de almacenamiento distribuido.

De acuerdo con esta realización de la presente invención, en la etapa 120, se puede llevar a cabo un cálculo de hash sobre un identificador de un bloque de datos o de un bloque de verificación de los M bloques de datos o de los N bloques de verificación para generar un valor de clave; y se determina un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación en función del valor de clave, y se utiliza el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

Por ejemplo, en caso de que se utilice el primer nodo de almacenamiento físico como un nodo de almacenamiento correspondiente a un primer bloque de un segmento de datos, los nodos de almacenamiento físicos correspondientes a otros bloques del segmento de datos se pueden determinar secuencialmente de acuerdo con números de serie del nodo de almacenamiento físico. Suponiendo que el número de serie del primer nodo de almacenamiento físico es 101, los números de serie de los nodos de almacenamiento físicos correspondientes a otros bloques del segmento de datos pueden ser 102, 103, 104, y similares.

Alternativamente, como otra realización, en la etapa 120, el cálculo de hash se puede llevar a cabo sobre un identificador de un segmento de datos que está dividido en M bloques de datos, de tal modo que se genera un valor de clave; y se determina un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación de acuerdo con el valor de clave, y se utiliza el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

En otras palabras, el valor de clave se puede calcular de acuerdo con el identificador del segmento de datos o el identificador de cualquier bloque de datos (por ejemplo, el primer bloque de datos o bloque de verificación) en el segmento de datos.

Por ejemplo, el cálculo de hash se puede llevar a cabo sobre un identificador de cada segmento de datos para obtener un valor de clave del segmento de datos, se busca en la tabla de hash distribuida un número de serie del nodo de almacenamiento físico correspondiente al valor de clave, y un bloque de los M bloques de datos y N bloques de verificación en el segmento de datos se habilita en correspondencia con el número de serie del nodo de almacenamiento físico. Esta realización de la presente invención no se limita a esto. El número de serie del nodo de almacenamiento físico correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación en el segmento de datos se puede determinar asimismo utilizando la tabla de hash distribuida y de acuerdo con un identificador de un determinado bloque de datos o bloque de verificación en cada segmento de datos, u otra información que puede identificar el segmento de datos.

Por ejemplo, el número de serie del nodo de almacenamiento físico determinado se puede utilizar como el número de serie del nodo de almacenamiento físico correspondiente al primer bloque de datos o bloque de verificación en el segmento de datos, y los números de serie de nodos de almacenamiento físicos correspondientes a otros bloques de datos o bloques de verificación en el segmento de datos se pueden determinar de acuerdo con un modo de ordenamiento (por ejemplo, un modo de ordenamiento en orden ascendente o descendente) que sea conforme con la regla preestablecida. El modo de ordenamiento basado en reglas no se limita a esta realización de la presente invención, siempre que se desplieguen por lo menos M+1 bloques en diferentes nodos de almacenamiento físicos. El despliegue se puede llevar a cabo en una secuencia global o una secuencia parcial, por ejemplo, y se puede llevar a cabo asimismo en un modo de ordenamiento entrelazado, un modo de ordenamiento de secuencia segmentada u otro modo de ordenamiento de secuencia a un intervalo fijo.

De acuerdo con esta realización de la presente invención, en la etapa 120, se determinan M+N diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, y en la etapa 130, los M bloques de datos y los N bloques de verificación se pueden almacenar independientemente en los M+N diferentes nodos de almacenamiento físicos.

De acuerdo con esta realización de la presente invención, se despliegan M+N bloques de datos en los M+N nodos de almacenamiento físicos, y en este caso, se puede garantizar que los segmentos de datos del archivo de datos no se pierden cuando fallan al mismo tiempo N nodos de almacenamiento físicos.

De acuerdo con esta realización de la presente invención, en la etapa 120, el primer nodo de almacenamiento físico puede corresponder a un bloque de los por lo menos M+1 bloques, y se determinan nodos de almacenamiento físicos correspondientes a otros por lo menos M bloques de los por lo menos M+1 bloques en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, donde los M+1 diferentes nodos de almacenamiento físicos incluyen el primer nodo de almacenamiento físico.

Por ejemplo, el primer nodo de almacenamiento físico puede ser el primer bloque, el bloque intermedio, el último bloqueo o cualquier bloque de los M+1 bloques. Utilizando como ejemplo el primer bloque de los M+1 bloques y suponiendo que el número de serie del primer nodo de almacenamiento físico es 100, los números de serie de los otros M bloques son 101, 102, ... y 100+M.

Alternativamente, como otra realización, se pueden determinar asimismo nodos de almacenamiento físicos correspondientes a los por lo menos M+1 bloques en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, donde M+1 diferentes nodos de almacenamiento físicos no incluyen el primer nodo de almacenamiento físico.

Por ejemplo, el primer nodo de almacenamiento físico es 100, y los M+1 nodos son 110, 111, ... y 110+M.

De acuerdo con esta realización de la presente invención, los diferentes nodos de almacenamiento físicos son discos duros, y el número de discos duros es mayor o igual que M+1.

En este caso, se puede garantizar que se despliegan M+N bloques de datos y bloques de verificación en diferentes discos duros, garantizando de ese modo que un fallo de cualquier disco duro no provoca una pérdida de segmentos de datos.

Alternativamente, como otra realización, los diferentes nodos de almacenamiento físicos son servidores, donde el número de servidores es mayor o igual que M+1.

En este caso, se puede garantizar que se despliegan M+N bloques de datos y bloques de verificación en discos duros en servidores diferentes, garantizando de ese modo que un fallo de cualquier servidor no provoca una pérdida de un segmento de datos.

Alternativamente, como otra realización, los diferentes nodos de almacenamiento físicos son bastidores, donde el número de bastidores es mayor o igual que M+1.

En este caso, se puede garantizar que se despliegan M+N bloques de datos y bloques de verificación en discos duros en servidores en bastidores diferentes, garantizando de ese modo que un fallo en cualquier bastidor no provoca una pérdida de segmentos de datos.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: en caso de que uno de los por lo menos M+1 diferentes nodos de almacenamiento físicos que almacenan por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación falle, restablecer datos que se encuentran en el nodo de almacenamiento físico que falla de los por lo menos M+1 diferentes nodos de almacenamiento físicos en un nodo de almacenamiento respaldo en caliente en un sistema de almacenamiento distribuido en el que está situado el nodo de almacenamiento físico.

Por ejemplo, el número de serie del nodo de almacenamiento físico que falla puede ser sustituido con el número de serie del nodo de almacenamiento físico de respaldo en caliente, de tal modo que se migran todos los bloques de datos o los bloques de verificación del nodo de almacenamiento físico que falla al nodo de almacenamiento físico de respaldo en caliente.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle y si L no es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es opuesto al sentido de una secuencia en el primer modo de ordenamiento, y si L es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es el mismo que el sentido de la secuencia en el primer modo de ordenamiento.

En otras palabras, cuando un nodo de almacenamiento físico falla, se migran solamente no más de (M+N)/2 bloques de datos y bloques de verificación para cada segmento de datos afectado. Por lo tanto, se reduce el impacto sobre otros nodos de almacenamiento físicos normales, reduciendo de ese modo la sobrecarga del sistema de almacenamiento distribuido para llevar a cabo un cálculo para una recuperación de fallos, mejorando la velocidad de la migración de datos y reduciendo el tiempo para la recuperación de fallos.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: en caso de que el nodo de almacenamiento físico del bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en el primer sentido.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: en caso de que el nodo de almacenamiento físico del bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de

los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en el segundo sentido.

De acuerdo con esta realización de la presente invención, se lleva a cabo una recuperación de fallos migrando, hacia delante o hacia atrás, bloques de datos o bloques de verificación almacenados en un nodo de almacenamiento físico que falla, de tal modo que los bloques de datos o bloques de verificación de un segmento de datos pueden seguir estando desplegados en nodos de almacenamiento físicos diferentes, garantizando de ese modo la fiabilidad del sistema de almacenamiento distribuido después de la recuperación de fallos.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos y si P no es mayor que (M+N)/2, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es el mismo que el sentido de la secuencia en el primer modo de ordenamiento; y si P es mayor que (M+N)/2, migrar un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es opuesto al sentido de la secuencia en el primer modo de ordenamiento, y P es un entero.

Dado que solamente se migran no más de (M+N)/2 bloques de datos y bloques de verificación en cada segmento de datos afectado, se puede mejorar la velocidad de la migración de datos, reduciendo de ese modo el tiempo de un proceso de expansión de la capacidad.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: en caso de que se añada un nuevo nodo de almacenamiento físico después del nodo de almacenamiento físico del bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en el primer sentido.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: en caso de que se añada un nuevo nodo de almacenamiento físico después del nodo de almacenamiento físico en el que está situado bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en el segundo sentido.

De acuerdo con esta realización de la presente invención, se lleva a cabo una expansión de la capacidad migrando, hacia delante o hacia atrás, bloques de datos o bloques de verificación almacenados en nodos de almacenamiento físicos, de tal modo que los bloques de datos o bloques de verificación de un segmento de datos pueden seguir estando desplegados en diferentes nodos de almacenamiento físicos, garantizando de ese modo la fiabilidad del sistema de almacenamiento distribuido después de la expansión de la capacidad.

Alternativamente, como otra realización, el procedimiento de la figura 1 incluye además: cuando es necesario leer el archivo de datos, determinar, utilizando el algoritmo aleatorio, el primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación; determinar por lo menos M nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento; leer por lo menos M bloques desde los por lo menos M diferentes nodos de almacenamiento físicos, donde los M bloques leídos incluyen los M bloques de datos o incluyen algunos bloques de datos de los M bloques de datos y algunos bloques de verificación de los N bloques de verificación; y llevar a cabo una descodificación y una verificación inversa sobre los por lo menos M bloques para generar M bloques de datos descodificados, y combinar los M bloques de datos descodificados para generar el archivo de datos.

De acuerdo con esta realización de la presente invención, el procedimiento para leer el archivo de datos exactamente el inverso al procedimiento para almacenar el archivo de datos. Por ejemplo, el sistema de almacenamiento distribuido puede recibir una solicitud de lectura enviada por el usuario a través del cliente,

autenticar al usuario de acuerdo con la información de identidad del usuario, donde la información de identidad está contenida en la solicitud de lectura, y permitir al usuario leer el archivo de datos solamente después de una identificación satisfactoria. Esto no se limita a esta realización de la presente invención. Por ejemplo, el usuario puede leer directamente el archivo de datos sin necesidad de autenticación.

El sistema de almacenamiento distribuido de esta realización de la presente invención puede asignar un identificador para el archivo de datos a leer, y en caso de que la solicitud de lectura incluya la información de usuario y el tipo de servicios de almacenamiento, puede asimismo asignar un identificador para el archivo de datos de acuerdo con la información de usuario y con el tipo de servicio de almacenamiento.

El sistema de almacenamiento distribuido en esta realización de la presente invención puede determinar, utilizando la tabla de hash distribuida y de acuerdo con un identificador asignado para cada segmento de datos del archivo de datos cuando el archivo de datos es almacenado, un nodo de almacenamiento físico correspondiente a cada segmento de datos.

El sistema de almacenamiento distribuido de esta realización de la presente invención puede utilizar un nodo de almacenamiento físico determinado como un nodo de almacenamiento físico correspondiente a un determinado bloque de datos o bloque de verificación de un segmento de datos, y se pueden determinar de acuerdo con una regla ascendente números de serie de nodos de almacenamiento físicos correspondientes a otros bloques de datos o bloques de verificación en el segmento de datos.

De acuerdo con esta realización de la presente invención, el archivo de datos se puede obtener desde el exterior del sistema de almacenamiento distribuido.

Alternativamente, como otra realización, el archivo de datos se puede obtener a partir de algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido.

De acuerdo con esta realización de la presente invención, el algoritmo de redundancia es el algoritmo de codificación de borrado, y el algoritmo aleatorio es el algoritmo de tabla de hash distribuida.

De acuerdo con esta realización de la presente invención, los nodos de almacenamiento físicos diferentes son nodos de almacenamiento físicos diferentes en el sistema de almacenamiento distribuido, cada nodo de almacenamiento físico de los nodos de almacenamiento físicos diferentes incluye múltiples nodos de almacenamiento virtuales, y se despliegan nodos de almacenamiento virtuales con números de serie consecutivos en los diferentes nodos de almacenamiento físicos de acuerdo con un segundo modo de ordenamiento que es conforme con la regla preestablecida.

La determinación, utilizando el algoritmo aleatorio, del primer nodo de almacenamiento físico correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación, y la determinación de los por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento incluye: determinar, utilizando el algoritmo aleatorio, un número de serie de un primer nodo de almacenamiento virtual correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación, y determinar números de serie de por lo menos M+1 nodos de almacenamiento virtuales en base al número de serie del primer nodo de almacenamiento virtual y de acuerdo con el primer modo de ordenamiento; y el almacenamiento de los por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos incluye: almacenar los por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en nodos de almacenamiento virtuales correspondientes a los números de serie de los por lo menos M+1 nodos de almacenamiento virtuales.

Por ejemplo, cada disco duro en el nodo de almacenamiento físico se puede dividir en múltiples particiones, es decir, múltiples nodos virtuales, y T nodos virtuales con números de serie consecutivos en los múltiples nodos virtuales se pueden asimismo desplegar en diferentes discos duros del sistema de almacenamiento distribuido de acuerdo con el modo de ordenamiento basado en reglas.

El segundo modo de ordenamiento que es conforme con la regla preestablecida puede ser igual o diferente al primer modo de ordenamiento que es conforme con la regla preestablecida, y en caso de que el segundo modo de ordenamiento sea diferente al primer modo de ordenamiento, es necesario que el primer modo de ordenamiento y el segundo modo de ordenamiento cumplan conjuntamente la condición de que se desplieguen bloques de datos o bloques de verificación en nodos de almacenamiento físicos diferentes del sistema de almacenamiento distribuido. Por ejemplo, el primer modo de ordenamiento es el modo de ordenamiento entrelazado, y el segundo modo de ordenamiento es el modo de ordenamiento de secuencia. En otras palabras, los números de serie de los nodos virtuales corresponden a números de serie de nodos de almacenamiento físicos diferentes utilizando un procedimiento de ordenamiento de secuencias, y los números de serie de los bloques de datos o de los bloques de verificación corresponden a los números de serie de los nodos virtuales utilizando un procedimiento de ordenamiento entrelazado. Esto puede garantizar asimismo que los bloques de datos y los bloques de verificación se despliegan en nodos de almacenamiento físicos lo más diferentes posible.

A diferencia de un procedimiento convencional para distribuir los mínimos bloques de datos utilizando la tabla de hash distribuida con el fin de implementar una distribución aleatoria de bloques de datos, en esta realización de la presente invención se adopta un procedimiento para distribuir datos con dos etapas: en la primera etapa, se lleva a cabo una distribución aleatoria sobre segmentos de datos (o sobre un determinado bloque de datos de cada segmento de datos) utilizando la tabla de hash distribuida; en la segunda etapa, se almacenan bloques de datos y bloques de verificación incluidos en cada segmento de datos en un modo de despliegue basado en reglas, donde el modo de despliegue basado en reglas puede ser un despliegue de secuencia global, un despliegue de secuencia parcial, un despliegue entrelazado o un despliegue de secuencia segmentada, y se puede utilizar un principio de despliegue con un algoritmo de ordenamiento de nodos virtuales en un disco duro, de tal modo que los bloques de datos y los bloques de verificación de cada segmento de datos se despliegan en nodos de almacenamiento físicos diferentes. Los nodos de almacenamiento físicos acordes con esta realización de la presente invención se pueden definir como discos duros, servidores o bastidores en función de los requisitos del usuario, con el fin de evitar que los datos en el sistema de almacenamiento distribuido no sean válidos debido a un fallo de un disco duro, un servidor o un bastidor.

A continuación se describen en detalle realizaciones de la presente invención haciendo referencia a ejemplos específicos.

La figura 2 es un diagrama esquemático de un proceso en el que un sistema de almacenamiento distribuido divide y almacena un archivo de datos, según una realización de la presente invención. Para mayor claridad, la figura 2 muestra solamente 5 segmentos de datos y muestra solamente 6 bloques de datos, 3 bloques de verificación y 9 nodos virtuales para un segmento de datos.

210: dividir un archivo de datos para generar K (por ejemplo, 5) segmentos de datos.

220: dividir adicionalmente un segmento de datos con una granularidad menor para generar M (por ejemplo, 6) bloques de datos.

230: añadir N (por ejemplo, 3) bloques de verificación utilizando un algoritmo de codificación con redundancia.

240: obtener, utilizando un resultado de un cálculo de hash, un número de serie de un nodo virtual correspondiente al segmento de datos, obtener, en base al número de serie del nodo virtual y de acuerdo con un modo de ordenamiento basado en reglas, números de serie de nodos virtuales correspondientes a los M bloques de datos y los N bloques de verificación, y finalmente almacenar secuencialmente los M bloques de datos y los N bloques de verificación en M+N (por ejemplo, 9) nodos virtuales VN.

Se debe observar que, como otra realización, un nodo virtual de esta realización puede ser sustituido por un nodo de almacenamiento físico, y el número de serie del nodo virtual puede ser sustituido por el número de serie del nodo de almacenamiento físico. No se dan a conocer detalles adicionales en la presente memoria.

La figura 3 es un diagrama de flujo esquemático de un proceso de almacenamiento distribuido, según una realización de la presente invención. El procedimiento de la figura 3 puede ser ejecutado mediante un sistema de almacenamiento distribuido. El procedimiento de la figura 3 puede incluir el siguiente contenido:

De acuerdo con esta realización de la presente invención, durante el establecimiento y la distribución de nodos virtuales (es decir, particiones) del sistema de almacenamiento distribuido, la distribución de los nodos virtuales puede satisfacer la condición de que no haya nodos virtuales consecutivos en un mismo nodo de almacenamiento físico, por ejemplo, las particiones Partición (k), Partición (k-1) y Partición (k+1) se despliegan en tres diferentes discos duros, servidores o bastidores.

310. Recibir una solicitud de almacenamiento de un usuario.

Por ejemplo, el sistema de almacenamiento distribuido puede recibir una solicitud de lectura enviada por el usuario a través de un cliente. La solicitud de lectura puede transportar información de identidad del usuario.

320. Autenticar la validez del almacenamiento del usuario, identificar un tipo de servicio, añadir un identificador de archivo y recibir un archivo de datos a almacenar.

Por ejemplo, si se autentica primero que el usuario sea un usuario autorizado, y después de que se autentique que el usuario es un usuario autorizado, se asigna un identificador del archivo de datos y se recibe el archivo de datos a almacenar. Por ejemplo, el identificador del archivo de datos puede ser "nombre de archivo + información del autor + información de tiempo + número de serie de versión".

330. Dividir cada archivo de datos para generar segmentos de datos, y añadir un identificador para cada segmento de datos.

Por ejemplo, un archivo de datos se puede dividir en K segmentos de datos, se añade un identificador para cada segmento de datos, y el identificador puede incluir el identificador del archivo de datos + el número de serie del segmento de datos.

340. Dividir cada segmento de datos para generar bloques de datos, y añadir un identificador para cada bloque de datos.

Por ejemplo, un segmento de datos se puede dividir en M bloques de datos, se añade un identificador para cada bloque de datos, por ejemplo, el identificador puede incluir un identificador del segmento de datos + el número de serie del bloque de datos.

350. Codificar contenido de bloques de datos de cada segmento de datos para generar bloques de verificación, y añadir identificadores.

Por ejemplo, se lleva a cabo una codificación de verificación cíclica sobre el contenido de M bloques de datos de un segmento de datos para generar N bloques de verificación, y se añade un identificador para cada bloque de verificación, por ejemplo, el identificador puede incluir el identificador del segmento de datos + el número de serie del bloque de verificación.

360. Realizar un cálculo de hash sobre el identificador de cada segmento de datos o un identificador de un determinado bloque de datos para obtener un valor de clave, y determinar, de acuerdo con el valor de clave, el número de serie de un nodo virtual correspondiente al segmento de datos.

El sistema de almacenamiento distribuido puede proporcionar externamente funciones de almacenamiento, lectura y eliminación en base a una tabla de hash distribuida. La tabla de hash distribuida puede incluir una relación de mapeo entre un valor de clave y un número de serie del nodo virtual. Cada pieza de datos almacenados o leídos tiene un identificador único en todo el grupo.

De acuerdo con esta realización de la presente invención, el cálculo de hash se puede llevar a cabo de acuerdo con un identificador de un segmento de datos para generar un valor de clave, y de acuerdo con el valor de clave se determina un número de serie de un nodo de almacenamiento virtual correspondiente al segmento de datos y una posición de despliegue de una unidad física en la que está situado el nodo virtual.

Alternativamente, de acuerdo con esta realización de la presente invención, el cálculo de hash se puede llevar a cabo asimismo de acuerdo con un identificador de un primer bloque de datos del segmento de datos o un identificador de cualquier bloque de datos del segmento de datos para generar un valor de clave, y se determina el número de serie de un nodo virtual correspondiente al segmento de datos buscando en la tabla de hash distribuida según el valor de clave.

370. Determinar, de acuerdo con el número de serie determinado del nodo de almacenamiento virtual y según un modo de ordenamiento basado en reglas, números de serie de M+N nodos virtuales necesarios para almacenar el segmento de datos.

Se obtienen números de serie de otros M+N-1 nodos virtuales, es decir, números de serie de N+M nodos virtuales, de acuerdo con el número de serie del nodo virtual obtenido en la etapa 360 y utilizando un algoritmo de ordenamiento basado en reglas (por ejemplo, ordenamiento de secuencia, ordenamiento de intervalo y secuencia segmentada). Utilizando como ejemplo el despliegue de secuencia, se despliegan M+N bloques de datos y bloques de verificación en M+N nodos virtuales consecutivos, y los M+N nodos virtuales consecutivos se despliegan en nodos de almacenamiento físicos independientes entre sí. Se supone que el primer bloque de un segmento de datos corresponde al número de serie del nodo virtual 100, el segundo bloque corresponde a un número de serie de nodo virtual 101, y así sucesivamente.

El nodo virtual cuyo número de serie es 100 y el nodo virtual cuyo número de serie es 101 se distribuyen en nodos de almacenamiento físicos diferentes, y por lo tanto el primer bloque y el segundo bloque del segmento de datos se pueden almacenar en nodos de almacenamiento físicos diferentes.

380. Almacenar bloques de datos y bloques de verificación de cada segmento de datos del archivo de datos en los nodos virtuales que se han determinado utilizando el cálculo de hash y el modo de ordenamiento basado en reglas.

Por ejemplo, un identificador de un segmento de datos de un archivo a almacenar puede ser una cadena de caracteres, y se pueden obtener valores 1-10000 (es decir, valores de clave) después de que se realice un cálculo de hash. Utilizando 100 como intervalo, 1-100 se sitúan en la partición A, 100-200 están en la siguiente partición B y 200-300 en otra partición C. Por ejemplo, cuando un valor de clave de un identificador de un segmento de datos es 1, 2, 50 ó 99, se selecciona una partición A como posición de almacenamiento de un bloque (por ejemplo, el primer bloque) del segmento de datos, y otros bloques de datos se almacenan en otros discos duros diferentes de acuerdo con el modo de ordenamiento basado en reglas definido en esta realización de la presente invención.

De acuerdo con esta realización de la presente invención, todos los nodos virtuales del sistema de almacenamiento distribuido pueden componer un anillo lógico. Por ejemplo, un sistema tiene 100 discos, cada disco tiene 100 nodos virtuales (particiones), y en este caso 10.000 nodos virtuales forman un anillo lógico. Los nodos virtuales, es decir, los números de serie de los nodos virtuales, distribuidos en el disco duro 1 son 1, 101, 201, 301, 401, 501, 601, 701, 801, 901, y así sucesivamente. En esta realización de la presente invención, se pueden desplegar secuencialmente N+M bloques de datos en los 10.000 nodos virtuales.

Alternativamente, en esta realización de la presente invención, un despliegue consecutivo se puede asimismo llevar a cabo parcialmente. Por ejemplo, se utilizan 50 nodos virtuales como un segmento parcial. Los N+M bloques de datos y bloques de verificación se despliegan secuencialmente en los 50 nodos virtuales. Suponiendo que 12 + 3 bloques de datos y bloques de verificación se inician a partir de la partición cuyo número de serie es 45, los bloques de datos y los bloques de verificación se pueden desplegar secuencialmente en nodos virtuales cuyos números de serie son 45, 46, 47, 48, 49, 50, 1, 2, 3, 4, 5, 6, 7, 8 y 9.

390. Devolver una respuesta de almacenamiento completo al cliente después de confirmar que el almacenamiento se ha completado.

De acuerdo con esta realización de la presente invención, cuando se distribuyen datos en nodos de almacenamiento físicos, se adopta un procedimiento de mapeo con dos etapas, es decir, un modo "aleatorio + ordenado" en el que un segmento de datos se despliega aleatoriamente en la primera etapa, y se despliegan bloques de datos de acuerdo con una regla en la segunda etapa. En esta realización de la presente invención, se puede garantizar que los bloques de datos son almacenados (o desplegados) en discos duros diferentes. Por lo tanto, los datos del segmento de datos siguen existiendo o pueden ser restablecidos cuando algún disco duro no es válido, evitando de ese modo una pérdida de datos provocada por un fallo de un disco de una sola unidad física, y mejorando sensiblemente la fiabilidad del almacenamiento de datos.

De acuerdo con esta realización de la presente invención, el modo de ordenamiento basado en reglas adoptado durante el almacenamiento de bloques de datos puede coincidir con el modo de ordenamiento basado en reglas de los nodos virtuales, por ejemplo, se distribuyen nodos virtuales consecutivos en discos duros diferentes, y los bloques de datos almacenados en los nodos virtuales consecutivos pueden garantizar que los bloques de datos se almacenan en discos duros diferentes. Si los nodos virtuales se distribuyen en discos duros diferentes utilizando un algoritmo de salto o de entrelazado, se puede asimismo ajustar correspondientemente una regla de distribución de los bloques de datos, pero los bloques de datos de un mismo segmento de datos siguen siempre estando distribuidos en discos duros diferentes. Se comprenderá que cualesquiera modos de ordenamiento basados en reglas que se utilicen conjuntamente para permitir que los bloques de datos se distribuyan en discos duros diferentes caerán dentro del alcance de protección de la presente invención.

Cabe señalar que, como otra realización, un nodo virtual de la realización de la figura 3 puede ser sustituido por un nodo de almacenamiento físico, y el número de serie del nodo virtual puede ser sustituido por el número de serie del nodo de almacenamiento físico. No se dan a conocer detalles adicionales en la presente memoria.

La figura 4 es un diagrama de flujo esquemático de un proceso de almacenamiento distribuido, según una realización de la presente invención. El procedimiento de la figura 4 puede ser ejecutado mediante un sistema de almacenamiento distribuido. El procedimiento de lectura de la figura 4 corresponde al procedimiento de almacenamiento de la figura 3, y algunas descripciones detalladas no se proporcionan en este caso. El procedimiento de la figura 4 incluye el siguiente contenido:

410. Recibir una solicitud de lectura de un usuario.

Por ejemplo, el sistema de almacenamiento distribuido puede recibir una solicitud de lectura enviada por el usuario a través de un cliente. La solicitud de lectura puede transportar información de identidad del usuario.

420. Autenticar la validez de lectura del usuario, identificar un tipo de servicio, y añadir un identificador de un archivo de datos.

Por ejemplo, se autentica primero que el usuario sea un usuario autorizado, y después de que sea autenticado que el usuario es un usuario autorizado, se determina un identificador de un archivo de datos a leer de acuerdo con un tipo de servicio de almacenamiento y con información del cliente, por ejemplo, el identificador del archivo de datos puede ser "nombre de archivo + información del autor + información del tiempo + número de serie de versión". El sistema de almacenamiento distribuido puede determinar identificadores de segmentos de datos del archivo de datos, de acuerdo con un registro que se genera cuando el archivo de datos es almacenado.

430. Determinar posiciones de almacenamiento de los segmentos de datos, los bloques de datos y los bloques de verificación del archivo de datos, de acuerdo con un cálculo de hash y con un modo de ordenamiento basado en reglas.

Por ejemplo, el cálculo de hash se lleva a cabo de acuerdo con un identificador determinado de cada segmento de datos para generar un valor de clave, y se determina un número de serie de un nodo virtual del segmento de datos de acuerdo con el valor de clave. Por consiguiente, el cálculo de hash puede asimismo llevarse a cabo de acuerdo con un identificador de un primer bloque de datos del segmento de datos o con un identificador de un determinado bloque de datos para generar un valor de clave, y se determina el número de serie de un nodo virtual del segmento de datos de acuerdo con el valor de clave, es decir, se determina el número de serie de un nodo virtual correspondiente a un determinado bloque de datos (por ejemplo, el primer bloque de datos) del segmento de datos. Se obtienen números de serie de otros M+N-1 nodos virtuales, es decir, números de serie de M+N nodos virtuales en total, de acuerdo con el número de serie del nodo virtual (es decir, una posición del nodo virtual) y utilizando un

algoritmo de ordenamiento basado en reglas (por ejemplo, ordenamiento de secuencia, ordenamiento de intervalo y secuencia segmentada).

440. Adquirir bloques de datos y bloques de verificación de cada segmento de datos.

Se leen los bloques de datos y los bloques de verificación, de acuerdo con los números de serie obtenidos de los M+N nodos virtuales, a partir de los nodos virtuales correspondientes a los números de serie de los nodos virtuales.

450. Llevar a cabo codificación y verificación inversa sobre los bloques de datos y bloques de verificación leídos, para obtener bloques de datos, y combinar los bloques de datos para generar segmentos de datos.

460. Combinar los segmentos de datos para obtener el archivo de datos.

470. Devolver una respuesta de lectura completa al cliente después de confirmar que la lectura del archivo de datos se ha completado.

La figura 5A es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos cuando un disco duro de un sistema de almacenamiento distribuido falla, según una realización de la presente invención. La figura 5B es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos de un sistema de almacenamiento distribuido después de una recuperación de fallos, según una realización de la presente invención. La figura 5C es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos de un sistema de almacenamiento distribuido, después de una recuperación de fallos, según otra realización de la presente invención. A continuación se utiliza un nodo de almacenamiento físico que es un disco duro, como un ejemplo para la descripción.

Haciendo referencia a la figura 5A, cada segmento de datos tiene 6 bloques de datos y 3 bloques de verificación, es decir, 9 bloques en total, y los 9 bloques están numerados secuencialmente. Un bloque cuyo número de serie es 1 del segmento de datos 1 está desplegado en el disco duro i, un bloque cuyo número de serie es 2 del segmento de datos 2 está desplegado en el disco duro i y así sucesivamente. Un bloque cuyo número de serie es 9 del segmento de datos 9 está desplegado en el disco duro i, y otros bloques de cada segmento de datos están desplegados secuencialmente en diferentes discos duros por analogía. En esta realización, se supone que el disco duro i falla.

Haciendo referencia a la figura 5B, un determinado número de discos duros de respaldo en caliente están reservados en el sistema de almacenamiento distribuido, y cuando un disco duro falla, se adopta un disco duro de respaldo en caliente para reemplazar directamente el número de serie del disco que falla. Por ejemplo, cuando el disco duro i falla, los datos que se encuentran en el disco duro i que falla se restablecen en un nuevo disco duro i'. El número de discos duros de respaldo en caliente se puede determinar de acuerdo con la fiabilidad del disco duro, por ejemplo, un mayor requisito de fiabilidad requiere un mayor número de discos duros de respaldo en caliente.

Haciendo referencia a la figura 5C, cuando el disco duro i falla, un determinado bloque de datos de los segmentos de datos se despliega en el disco duro i, y los números de serie de los bloques de datos son 1 a 9; se lleva a cabo un restablecimiento de los datos adoptando el procedimiento de migrar secuencialmente hacia delante bloques cuyos números de serie no son mayores que 9/2 y migrar secuencialmente hacia atrás otros bloques de acuerdo con diferentes números de serie de los bloques desplegados en el disco duro i. En este caso, la cantidad de datos migrados es la mínima, y se puede seguir manteniendo el despliegue ordenado de los bloques de datos, consiguiendo de ese modo la mejor combinación de rendimiento y fiabilidad. A continuación se utiliza M+N adoptando una protección de codificación 6 + 3, como ejemplo para la descripción. Se ajusta una regla de migración de datos reales hacia delante y hacia atrás en función del número total de bloques, siguiendo el principio de satisfacer la condición de que la cantidad de datos migrados sea mínima.

Para el segmento de datos 1, cuando se pierde el bloque cuyo número de serie es 1, el bloque cuyo número de serie es 1 se puede calcular y restablecer de acuerdo con los otros 8 bloques, y el bloque cuyo número de serie es 1 se puede restablecer en el disco duro i-1 debido a que 1 no es mayor que (6+3)/2; para el segmento de datos 2, cuando se pierde el bloque cuyo número de serie es 2, el bloque cuyo número de serie es 2 se puede calcular y restablecer de acuerdo con los otros 8 bloques, y debido a que 2 no es mayor que (6+3)/2, el bloque cuyo número de serie es 1 se puede migrar hacia delante desde el disco duro i-1, es decir, migrar al disco duro i-2, y el bloque cuyo número de serie es 2 se restablece en el disco duro i-1, y así sucesivamente. Además, para el segmento de datos 5, cuando se pierde el bloque cuyo número de serie es 5, el bloque cuyo número de serie es 5 se puede calcular y restablecer de acuerdo con otros 8 bloques, y dado que 5 es mayor que (6+3)/2, el bloque cuyo número de serie es 5 se puede restablecer en el disco duro i+1, y los bloques cuyos números de serie son 6, 7, 8 y 9 se migran hacia atrás secuencialmente desde el disco duro i+1, el disco duro i+2, el disco duro i+3 y el disco duro i+4, es decir, se migran al disco duro i+2, al disco duro i+3, al disco duro i+4 y al disco duro i+5, y así sucesivamente.

Se debe entender que en la realización específica anterior, se utiliza como ejemplo para la descripción un nodo de almacenamiento físico que es un disco duro, y el procedimiento de recuperación de fallos en la realización anterior aplica asimismo a un escenario en el que el nodo de almacenamiento físico es un servidor o un bastidor, y no se dan a conocer detalles adicionales en la presente memoria.

La figura 6A es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos cuando se lleva a cabo expansión de la capacidad para un sistema de almacenamiento distribuido, según una realización de la presente invención; la figura 6B es un diagrama esquemático de posiciones de almacenamiento de algunos segmentos de datos después de que se lleva a cabo expansión de la capacidad para un sistema de almacenamiento distribuido, según una realización de la presente invención.

Cuando se lleva a cabo expansión de la capacidad para el sistema de almacenamiento distribuido, la expansión de la capacidad se puede llevar a cabo adoptando un procedimiento opuesto al procedimiento de recuperación de fallos de la figura 5B si se añade solamente un número pequeño de discos duros, por ejemplo, el porcentaje del número de discos duros añadidos en el número total de discos duros es menor que un umbral específico. Por ejemplo, se puede añadir el número de serie de un disco duro, y algunos bloques de datos y/o bloques de verificación de un disco duro adyacente se migran a los discos duros recién añadidos adoptando un procedimiento de migración hacia delante o hacia atrás, con el fin de garantizar un despliegue basado en reglas de los bloques de datos y los bloques de verificación.

Por ejemplo, haciendo referencia a la figura 6A, cada segmento de datos tiene 6 bloques de datos y 3 bloques de verificación, es decir, 9 bloques en total, y los 9 bloques están numerados secuencialmente. El bloque cuyo número de serie es 1 del segmento de datos 1 está desplegado en el disco duro i, el bloque cuyo número de serie es 2 del segmento de datos 2 está desplegado en el disco duro i y así sucesivamente. El bloque cuyo número de serie es 9 del segmento de datos 9 está desplegado en el disco duro i, y otros bloques de cada segmento de datos están desplegados secuencialmente en diferentes discos duros por analogía. En esta realización, se supone que se añade el disco duro i' entre el disco duro i y el disco duro i+1 del sistema de almacenamiento distribuido.

Haciendo referencia a la figura 6B, cuando el disco duro i' se añade después de un bloque cuyo número de serie es mayor que 1 del segmento de datos 1, dado que 1 no es mayor que (6+3)/2, el bloque cuyo número de serie es 1 se migra hacia atrás desde el disco duro i, es decir, se migra al disco duro i'; cuando el disco duro i' se añade después de un bloque cuyo número de serie es 2 del segmento de datos 2, debido a que 2 no es mayor que (6+3)/2, el bloque cuyo número de serie es 2 se migra hacia atrás desde el disco duro i, es decir, se migra al disco duro i', y el bloque cuyo número de serie es 1 se migra hacia atrás desde el disco duro i-1, es decir, se migra al disco duro i, y así sucesivamente. Además, dado que el disco duro i' se añade después de un bloque cuyo número de serie es 5 del segmento de datos 5, debido a que 5 es mayor que (6+3)/2, los bloques cuyo número de serie son 6, 7, 8 y 9 del segmento de datos 5 se migran secuencialmente hacia delante, es decir, se migran desde el disco duro i+1, el disco duro i+2, el disco duro i+3 y el disco duro i+4 al disco duro i', el disco duro i+1, el disco duro i+2 y el disco duro i+3, y así sucesivamente.

Alternativamente, como otra realización, cuando se lleva a cabo expansión de la capacidad a gran escala para un sistema, por ejemplo, el porcentaje del número de discos duros añadidos en el número total de discos duros es mayor que un umbral específico, los discos duros añadidos se pueden combinar en un nuevo dominio de función de almacenamiento, y en el nuevo dominio de función de almacenamiento se lleva a cabo mapeo de datos adoptando un procedimiento similar al procedimiento de mapeo con dos etapas de la realización anterior. Esto divide de manera efectiva un sistema de almacenamiento distribuido en diferentes dominios de fallo además de garantizar que el procedimiento de mapeo con dos etapas sigue siendo válido, mejorando de ese modo la fiabilidad de los datos de todo el sistema.

Alternativamente, como otra realización, cuando se lleva a cabo una expansión de la capacidad a gran escala para un sistema, se pueden volver a desplegar nodos virtuales en todos los discos duros (incluyendo discos duros anteriores a la expansión de la capacidad y discos duros recién añadidos) de acuerdo con un modo de ordenamiento original basado en reglas, y se lleva a cabo la migración de datos necesaria, de tal modo que el sistema después de la expansión de la capacidad sigue satisfaciendo el requisito de ordenamiento basado en reglas anterior a la realización de la expansión de la capacidad para el sistema, reduciendo de ese modo la dificultad y la carga de trabajo de mantenimiento subsiguiente, y garantizando la fiabilidad de los datos del sistema. En este caso, el proceso de migración de datos es equivalente a adquirir datos desde un sistema de almacenamiento distribuido original (es decir, una parte de un nuevo sistema de almacenamiento distribuido posterior a la expansión de la capacidad) y volver a desplegar los datos en el nuevo sistema de almacenamiento distribuido de acuerdo con el procedimiento de esta realización de la presente invención.

Se debe entender que en la realización específica anterior, se utiliza como ejemplo para la descripción un nodo de almacenamiento físico que es un disco duro, y el procedimiento de expansión de la capacidad en la realización anterior aplica asimismo a un escenario en el que el nodo de almacenamiento físico es un servidor o un bastidor, y no se dan a conocer detalles adicionales en la presente memoria.

Lo anterior describe el procedimiento de almacenamiento distribuido de acuerdo con realizaciones de la presente invención, y a continuación se describe un aparato de almacenamiento distribuido de acuerdo con realizaciones de la presente invención, haciendo referencia a las figuras 7 a 12.

La figura 7 es un diagrama estructural esquemático de un aparato de almacenamiento distribuido 700, según una realización de la presente invención. El aparato de almacenamiento distribuido 700 incluye: un módulo de generación 710, un módulo de determinación 720 y un módulo de almacenamiento 730.

El módulo de generación 710 divide un archivo de datos para generar K segmentos de datos, divide cada segmento de datos de los K segmentos de datos para generar M bloques de datos para cada segmento de datos y lleva a cabo codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación. El módulo de determinación 720 determina, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determina por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida. El módulo de almacenamiento 730 almacena por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros.

En esta realización de la presente invención, se puede determinar un nodo de almacenamiento correspondiente a un bloque de datos del archivo de datos utilizando el algoritmo aleatorio, se determinan por lo menos M+1 diferentes nodos de almacenamiento en base al nodo de almacenamiento determinado y de acuerdo con un modo de ordenamiento basado en reglas, y se almacenan por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento, de tal modo que se pueden desplegar bloques de datos de segmentos de datos en nodos de almacenamiento lo más diferentes posible, de acuerdo con el modo de ordenamiento basado en reglas, y se reduce la pérdida de datos que puede resultar de un fallo de un punto de cantar, mejorando de ese modo la fiabilidad del sistema de almacenamiento distribuido.

Para las operaciones y funciones de cada módulo del aparato de almacenamiento distribuido 700, se puede hacer referencia a las etapas 110, 120 y 130 del procedimiento de la figura 1. No se dan a conocer en este caso detalles adicionales para evitar su repetición.

La figura 8 es un diagrama estructural esquemático de un aparato de almacenamiento distribuido 800, según una realización de la presente invención. El aparato de almacenamiento distribuido 800 incluye: un módulo de generación 810, un módulo de determinación 820 y un módulo de almacenamiento 830. El módulo de generación 810, el módulo de determinación 820 y el módulo de almacenamiento 830 son similares al módulo de generación 710, el módulo de determinación 720 y el módulo de almacenamiento 730 de la figura 7, y no se dan a conocer detalles adicionales en este caso.

De acuerdo con esta realización de la presente invención, el módulo de determinación 820 lleva a cabo un cálculo de hash sobre un identificador de un bloque de datos o de un bloque de verificación de M bloques de datos o N bloques de verificación para generar un valor de clave, determina, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utiliza el nodo de almacenamiento físico como un primer nodo de almacenamiento físico.

De acuerdo con esta realización de la presente invención, un primer modo de ordenamiento basado en reglas es un modo de ordenamiento de secuencia global, un modo de ordenamiento de secuencia parcial, un modo de ordenamiento entrelazado o un modo de ordenamiento de secuencia segmentada.

De acuerdo con esta realización de la presente invención, el módulo de determinación 820 está configurado para llevar a cabo un cálculo de hash sobre un identificador de un segmento de datos que está dividido en M bloques de datos, con el fin de generar un valor de clave, determinar, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utilizar el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

De acuerdo con esta realización de la presente invención, el módulo de determinación 820 determina M+N nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, y el módulo de almacenamiento 830 almacena por separado los M bloques de datos y los N bloques de verificación en los M+N nodos de almacenamiento físicos diferentes.

De acuerdo con esta realización de la presente invención, el módulo de determinación 820 mapea el primer nodo de almacenamiento físico a un bloque de los por lo menos M+1 bloques, y determina, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a otros por lo menos M bloques de los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos incluyen el primer nodo de almacenamiento físico.

Alternativamente, como otra realización, el módulo de determinación 820 determina, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a por lo menos M+1 bloques, donde M+1 diferentes nodos de almacenamiento físicos no incluyen el primer nodo de almacenamiento físico.

De acuerdo con esta realización de la presente invención, los diferentes nodos de almacenamiento físicos son discos duros, y el número de discos duros es mayor que M+1.

Alternativamente, como otra realización, los diferentes nodos de almacenamiento físicos son servidores, donde el número de servidores es mayor que M+1.

Alternativamente, como otra realización, los diferentes nodos de almacenamiento físicos son bastidores, donde el número de bastidores es mayor que M+1.

Alternativamente, como otra realización, el aparato de almacenamiento distribuido 800 incluye además: un módulo de restablecimiento 850. El módulo de restablecimiento 850 está configurado para, en caso de que uno de por lo menos M+1 diferentes nodos de almacenamiento físicos que almacenan por lo menos M+1 bloques de M bloques de datos y N bloques de verificación falle, restablecer datos del nodo de almacenamiento físico que falla de los por lo menos M+1 diferentes nodos de almacenamiento físicos en un nodo de almacenamiento de respaldo en caliente en un sistema de almacenamiento distribuido en el que está situado el nodo de almacenamiento físico.

Alternativamente, como otra realización, el aparato de almacenamiento distribuido 800 incluye además: un módulo de restablecimiento 850.

El módulo de restablecimiento 850 está configurado para, en caso de que falle un nodo de almacenamiento físico de un bloque L-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos y si L no es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es opuesto al sentido de una secuencia en el primer modo de ordenamiento, y si L es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es el mismo que el sentido de la secuencia en el primer modo de ordenamiento.

Alternativamente, como otra realización, el módulo de restablecimiento 850 migra secuencialmente, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos falle, el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido.

Alternativamente, como otra realización, el módulo de restablecimiento 850 migra secuencialmente, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos falle, el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Alternativamente, como otra realización, el aparato de almacenamiento distribuido 800 incluye además: un módulo de expansión de la capacidad 860.

El módulo de expansión de la capacidad 860 migra, en caso de que un nuevo nodo de almacenamiento físico se añada después de un nodo de almacenamiento físico de un bloque P-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos y si P no es mayor que (M+N)/2, el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migra secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es el mismo que el sentido de una secuencia en el primer modo de ordenamiento, y si P es mayor que (M+N)/2, migra un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migra secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es opuesto al sentido de la secuencia en el primer modo de ordenamiento, y P es un entero.

Alternativamente, como otra realización, el módulo de expansión de la capacidad 860 está configurado para, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques

anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido.

Alternativamente, como otra realización, el módulo de expansión de la capacidad 860 migra, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico en el que está localizado un bloque P-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos, un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migra secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Alternativamente, como otra realización, el módulo de determinación 820 está configurado además para, cuando es necesario leer un archivo de datos, determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de M bloques de datos o N bloques de verificación, y determinar por lo menos M diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico determinado y de acuerdo con un primer modo de ordenamiento, donde el aparato de almacenamiento distribuido 800 incluye además: un módulo de lectura 840. El módulo de lectura 840 está configurado para leer por lo menos M bloques a partir de los M diferentes nodos de almacenamiento físicos, donde los M bloques leídos incluyen los M bloques de datos o incluyen algunos bloques de datos de los M bloques de datos y algunos bloques de verificación de los N bloques de verificación; y el módulo de generación 810 está configurado además para llevar a cabo descodificación y verificación inversa sobre los por lo menos M bloques para generar M bloques de datos descodificados, y combinar los M bloques de datos descodificados para generar el archivo de datos.

Alternativamente, como otra realización, el sistema de almacenamiento distribuido incluye además: un módulo de adquisición 870 configurado para adquirir un archivo de datos desde fuera del sistema de almacenamiento distribuido, o adquirir el archivo de datos desde algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido.

De acuerdo con esta realización de la presente invención, un algoritmo de redundancia es un algoritmo de codificación de borrado, y un algoritmo aleatorio es un algoritmo de tabla de hash distribuida.

De acuerdo con esta realización de la presente invención, los nodos de almacenamiento físicos diferentes son nodos de almacenamiento físicos diferentes en el sistema de almacenamiento distribuido, cada nodo de almacenamiento físico de los nodos de almacenamiento físicos diferentes incluye múltiples nodos de almacenamiento virtuales, y se despliegan nodos de almacenamiento virtuales con números de serie consecutivos en los diferentes nodos de almacenamiento físicos de acuerdo con un segundo modo de ordenamiento que es conforme con una regla preestablecida. El módulo de determinación 820 determina, utilizando el algoritmo aleatorio, un número de serie de un primer nodo de almacenamiento virtual correspondiente a un bloque de M bloques de datos y N bloques de verificación, y determina números de serie de por lo menos M+1 nodos de almacenamiento virtuales en base al número de serie del primer nodo de almacenamiento virtual y de acuerdo con un primer modo de ordenamiento; y el módulo de almacenamiento 730 almacena por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en nodos de almacenamiento virtuales correspondientes a los números de serie de los por lo menos M+1 nodos de almacenamiento virtuales.

De acuerdo con esta realización de la presente invención, el primer modo de ordenamiento o el segundo modo de ordenamiento se refieren a un procedimiento de ordenamiento secuencial a un intervalo fijo.

Para las operaciones y funciones de cada módulo del aparato de almacenamiento distribuido 800, se puede hacer referencia a las etapas 110, 120 y 130 del procedimiento de la figura 1. No se dan a conocer en este caso detalles adicionales para evitar su repetición.

La figura 9 es un diagrama estructural esquemático de un sistema de almacenamiento distribuido 900, según una realización de la presente invención. El sistema de almacenamiento distribuido 900 incluye: un cliente 910, múltiples discos duros 920 y un aparato de almacenamiento distribuido 930. El aparato de almacenamiento distribuido 930 puede ser el aparato de almacenamiento distribuido 700 de la figura 7 y el aparato de almacenamiento distribuido 800 de la figura 8. No se dan a conocer detalles adicionales en la presente memoria.

El aparato de almacenamiento distribuido 930 almacena/lee un archivo de datos de un usuario en/desde los múltiples discos duros 920 de acuerdo con una solicitud de almacenamiento/lectura enviada por el usuario a través del cliente 910.

En esta realización de la presente invención, el número de serie de un nodo virtual correspondiente a un segmento de datos de un archivo de datos se puede determinar en primer lugar utilizando un cálculo de hash, los números de serie de nodos virtuales correspondientes a cada bloque de datos y bloque de verificación del segmento de datos se

determinan en base al número de serie del nodo virtual y de acuerdo con un modo de ordenamiento basado en reglas, y los bloques de datos y los bloques de verificación se almacenan en correspondientes nodos virtuales.

Los nodos virtuales consecutivos están distribuidos en nodos de almacenamiento físicos diferentes, de tal modo que los bloques de datos del segmento de datos se pueden desplegar en nodos de almacenamiento físicos lo más diferentes posible de acuerdo con un modo de ordenamiento basado en reglas, y se reduce la pérdida de datos que puede resultar de un fallo de un punto de cantar, mejorando de ese modo la fiabilidad del sistema de almacenamiento distribuido.

La figura 10 es un diagrama de arquitectura de un sistema de almacenamiento distribuido 1000, según una realización de la presente invención. El sistema de almacenamiento distribuido 1000 es un ejemplo de la realización de la figura 9. El sistema de almacenamiento distribuido 1000 incluye: un cliente 1010 y un sistema de servidores de almacenamiento distribuido 1020.

El cliente 1010 puede conectar con el sistema de servidores de almacenamiento 1020 a través de internet.

El cliente 1010 puede ejecutar un programa agente del cliente del sistema de almacenamiento distribuido, donde el programa agente del cliente está configurado para soportar aplicaciones de almacenamiento distribuido de diversos tipos al acceder al sistema de almacenamiento distribuido, por ejemplo, el programa agente de cliente puede implementar almacenamiento y respaldo en línea personal, almacenamiento y respaldo en línea de empresas, almacenamiento en línea de aplicaciones, otro almacenamiento y respaldo emergente, o similares.

El sistema de servidores de almacenamiento distribuido 1020 puede incluir: un servidor de control 1030, un servidor de operaciones, administración y mantenimiento (OAM, operation, administration and maintenance) 1040, un servidor de servicio 1050, un grupo de recursos de almacenamiento 1070 y un motor de almacenamiento 1080. El motor de almacenamiento 1080 puede ser un ejemplo del aparato de almacenamiento distribuido de la figura 8.

El servidor de control 1030 está configurado principalmente para controlar el sistema de almacenamiento distribuido con el fin de ejecutar diversos servicios de almacenamiento, por ejemplo, organizar migración de datos, migración, y cancelación en punto crítico de respaldo y almacenamiento

El servidor de operaciones, administración y mantenimiento 1040 puede proporcionar una interfaz de configuración y una interfaz de operaciones y mantenimiento de un sistema de almacenamiento, y proporcionar funciones tales como funciones de registro y alarma.

El servidor de servicio 1050 puede proporcionar funciones tales como identificación de servicio y autenticación, y completar una función de transferencia de servicio.

El grupo de recursos de almacenamiento 1070 puede incluir un grupo de recursos de almacenamiento formado por nodos de almacenamiento físicos, por ejemplo, puede estar formado por servidores/placas de almacenamiento 1060, nodos virtuales en cada nodo de almacenamiento físico componen un anillo de almacenamiento lógico, y un archivo de datos de un usuario puede ser almacenado en los nodos virtuales en el grupo de recursos de almacenamiento.

El motor de almacenamiento 1080 puede proporcionar lógica de funciones principales del sistema de almacenamiento distribuido, y la lógica se puede desplegar en un determinado dispositivo del servidor de control 1030, el servidor de servicio 1050 y el servidor de operaciones, administración y mantenimiento 1040 y se puede desplegar asimismo en el servidor de control 1040, el servidor de servicio 1050, el servidor de operaciones, administración y mantenimiento 1040 y el grupo de recursos de almacenamiento 1070 en un modo de despliegue distribuido.

La figura 11A es un diagrama estructural esquemático de un servidor de almacenamiento/placa de almacenamiento 1100 de un sistema de almacenamiento distribuido, según una realización de la presente invención. La figura 11B es un diagrama estructural esquemático de un disco duro de un sistema de almacenamiento distribuido, según una realización de la presente invención. La figura 11C es un diagrama esquemático de un anillo lógico de nodos virtuales de un sistema de almacenamiento distribuido, según una realización de la presente invención. El servidor de almacenamiento/placa de almacenamiento 1100 puede ser un ejemplo del servidor de almacenamiento/placa de almacenamiento 1060 de la figura 10.

Haciendo referencia a la figura 11A, el servidor de almacenamiento/placa de almacenamiento 1100 puede incluir múltiples discos duros 1110.

Haciendo referencia a la figura 11B, un disco duro 1110 es una unidad de almacenamiento físico, y se puede dividir en múltiples nodos virtuales o particiones VN1 a VNn. Si un disco duro falla, esto puede provocar una pérdida de datos de todos los nodos virtuales en el disco duro.

Haciendo referencia a la figura 11C, múltiples nodos virtuales o particiones VN1 a VNn forman un anillo lógico de nodos virtuales, es decir, VNn y VN son dos nodos virtuales consecutivos lógicamente. Cada nodo virtual de los nodos de almacenamiento virtuales VN1 a VNn puede almacenar una determinada cantidad de datos.

Menos datos en un disco duro en el sistema de almacenamiento distribuido tienen como resultado una mayor probabilidad de que los bloques de datos y bloques de verificación se distribuyan en un mismo disco duro, y por lo tanto tienen como resultado una probabilidad mayor de que se almacenen múltiples bloques en un mismo disco duro. Cuando el sistema de almacenamiento distribuido cuando el sistema de almacenamiento distribuido tiene solamente 50 discos duros, la probabilidad de que 4 bloques de 15 bloques (que incluyen 12 bloques de datos y 3 bloques de verificación) se almacenen en un mismo disco duro es sensiblemente mayor que la probabilidad de que 4 bloques de 15 bloques se almacenen en un mismo disco duro cuando el sistema de almacenamiento distribuido tiene 100 discos duros.

En esta realización de la presente invención se mejora significativamente la fidelidad de los datos, es decir, la fiabilidad del sistema de almacenamiento distribuido. Si se adopta el procedimiento de esta realización de la presente invención, la fiabilidad se mejora significativamente para un sistema de cualquier escala. En este caso, la fiabilidad de los datos almacenados es irrelevante para la escala de un sistema, y por lo tanto la fiabilidad cambia notablemente, especialmente para un sistema de almacenamiento a pequeña escala.

Haciendo referencia a la tabla 3, cuando el tiempo medio de reparación (Mean Time To Repair, MTTR) del sistema de almacenamiento distribuido es de 730 horas, la fiabilidad de los datos almacenados puede llegar a 0,98876658, y cuando el MTTR del sistema de almacenamiento distribuido es de 22 horas, la fiabilidad de los datos almacenados puede llegar a 0,999999553.

Tabla 3 Resultado de mejora de la fiabilidad producido por las tecnologías de la presente invención

Configuración del sistema de almacenamiento	**Fidelidad de los datos que puede ser implementada en la presente invención**	**Fidelidad de los datos en la técnica anterior**	**Resultado de mejora real**
MTTR = 22 horas, número de discos duros = 50	0,99999955	0,87946	99,9996%
MTTR = 22 horas, número de discos duros = 100	0,99999955	0,98183	99,9975%
MTTR = 22 horas, número de discos duros = 300	0,99999955	0,99916	99,9468%
MTTR = 730 horas, número de discos duros = 50	0,98876658	0,74304	95,6283%
MTTR = 730 horas, número de discos duros = 100	0,98876658	0,90758	87,8452%
MTTR = 730 horas, número de discos duros = 300	0,98876658	0,96935	63,3494%

Se debe observar que el valor de la fiabilidad de los datos máxima es 1, y se puede considerar que una fidelidad de los datos menor que 0,9 soporta difícilmente una aplicación comercial.

Para una solución de almacenamiento (que incluye sistemas de almacenamiento con una tecnología de codificación de borrado y otras tecnologías similares de respaldo de datos) que adopta una tecnología de codificación con redundancia de datos, el sistema de almacenamiento distribuido acorde con esta realización de la presente invención elimina un punto único de fallo y elimina el riesgo potencial de que un fallo de un único disco duro, servidor o bastidor (armario) de conjunto de discos duros (discos) provoque una pérdida de datos, mejorando de ese modo la fiabilidad de los datos del sistema de almacenamiento distribuido.

La figura 12 es un diagrama estructural esquemático de un aparato de almacenamiento distribuido, según otra realización de la presente invención.

Un procesador 1210 invoca, por medio de un bus de comunicaciones 1230, código almacenado en una memoria 1220, donde el código se utiliza para dividir un archivo de datos con el fin de generar K segmentos de datos, dividir cada segmento de datos de los K segmentos de datos para generar M bloques de datos para cada segmento de datos, y llevar a cabo una codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación; determina, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determina por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida; y almacena por lo menos M+1 bloques de los M bloques de datos y de los N

bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros.

En esta realización de la presente invención, se puede determinar un nodo de almacenamiento físico correspondiente a un bloque de datos del archivo de datos utilizando el algoritmo aleatorio, se determinan por lo menos M+1 diferentes nodos de almacenamiento en base al nodo de almacenamiento físico determinado y de acuerdo con un modo de ordenamiento basado en reglas, y se almacenan por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, de tal modo que los bloques de datos de los segmentos de datos se pueden desplegar en nodos de almacenamiento físicos lo más diferentes posible de acuerdo con el modo de ordenamiento basado en reglas, y se reduce la pérdida de datos que puede resultar de un fallo de un punto de cantar, mejorando de ese modo la fiabilidad de un sistema de almacenamiento distribuido.

De acuerdo con esta realización de la presente invención, el procesador 1210 lleva a cabo un cálculo de hash sobre un identificador de un bloque de datos o de un bloque de verificación de los M bloques de datos o de los N bloques de verificación para generar un valor de clave, determina, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utiliza el nodo de almacenamiento físico como un primer nodo de almacenamiento físico.

De acuerdo con esta realización de la presente invención, el primer modo de ordenamiento basado en reglas es un modo de ordenamiento de secuencia global, un modo de ordenamiento de secuencia parcial, un modo de ordenamiento entrelazado o un modo de ordenamiento de secuencia segmentada.

De acuerdo con esta realización de la presente invención, el procesador 1210 lleva a cabo un cálculo de hash sobre un identificador de un segmento de datos que está dividido en M bloques de datos, con el fin de generar un valor de clave, determina, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utiliza el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

De acuerdo con esta realización de la presente invención, el procesador 1210 determina M+N nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, y almacena los M bloques de datos y los N bloques de verificación en los M+N nodos de almacenamiento físicos diferentes.

De acuerdo con esta realización de la presente invención, el procesador 1210 mapea el primer nodo de almacenamiento físico a un bloque de los por lo menos M+1 bloques, y determina, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a otros por lo menos M bloques de los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos incluyen el primer nodo de almacenamiento físico; o determina, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos no incluyen el primer nodo de almacenamiento físico.

De acuerdo con esta realización de la presente invención, los diferentes nodos de almacenamiento físicos son discos duros, y el número de discos duros es mayor o igual que M+1.

Alternativamente, como otra realización, los diferentes nodos de almacenamiento físicos son servidores, donde el número de servidores es mayor o igual que M+1.

Alternativamente, como otra realización, los diferentes nodos de almacenamiento físicos son bastidores, donde el número de bastidores es mayor o igual que M+1.

Alternativamente, como otra realización, el procesador 1210 restablece además, en caso de que uno de los por lo menos M+1 diferentes nodos de almacenamiento físicos que almacenan por lo menos M+1 bloques de M bloques de datos y N bloques de verificación falle, datos del nodo de almacenamiento físico que falla de los por lo menos M+1 diferentes nodos de almacenamiento físicos en un nodo de almacenamiento de respaldo en caliente en un sistema de almacenamiento distribuido en el que está situado el nodo de almacenamiento físico.

Alternativamente, como otra realización, el procesador 1210 para migra además secuencialmente, en caso de que falle un nodo de almacenamiento físico de un bloque L-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos y si L no es mayor que (M+N)/2, el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es opuesto al sentido de una secuencia en el primer modo de ordenamiento, y si L es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de

almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es el mismo que el sentido de la secuencia en el primer modo de ordenamiento.

Alternativamente, como otra realización, el procesador 1210 migra secuencialmente, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos falle, el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido.

Alternativamente, como otra realización, el procesador 1210 migra secuencialmente, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos falle, el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Alternativamente, como otra realización, el procesador 1210 migra además, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos y si P no es mayor que (M+N)/2, el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migra secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es el mismo que el sentido de una secuencia en el primer modo de ordenamiento, y si P es mayor que (M+N)/2, migra un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migra secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es opuesto al sentido de la secuencia en el primer modo de ordenamiento, y P es un entero.

Alternativamente, como otra realización, el procesador 1210 migra, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos, el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migra secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido.

Alternativamente, como otra realización, el procesador 1210 migra, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico en el que está localizado un bloque P-ésimo de M bloques de datos y N bloques de verificación que están ordenados de acuerdo con un primer modo de ordenamiento y almacenados en por lo menos M+1 diferentes nodos de almacenamiento físicos, un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migra secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido.

Alternativamente, como otra realización, el procesador 1210 está configurado además para, cuando un archivo de datos tiene que ser leído, determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o un bloque de verificación de M bloques de datos o N bloques de verificación, y determinar por lo menos M nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico determinado y de acuerdo con un primer modo de ordenamiento; el procesador 1210 está configurado además para leer por lo menos M bloques a partir de los M diferentes nodos de almacenamiento físicos, donde los M bloques leídos incluyen los M bloques de datos o incluyen algunos bloques de datos de los M bloques de datos y algunos bloques de verificación de los N bloques de verificación; y el procesador 1210 está configurado además para llevar a cabo descodificación y verificación inversa de los por lo menos M bloques para generar M bloques de datos descodificados, y combinar los M bloques de datos descodificados para generar el archivo de datos.

Alternativamente, como otra realización, el sistema de almacenamiento distribuido incluye además: una interfaz de E/S 1240, configurada para adquirir un archivo de datos desde fuera del sistema de almacenamiento distribuido, o adquirir el archivo de datos desde algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido.

De acuerdo con esta realización de la presente invención, los nodos de almacenamiento físicos diferentes son nodos de almacenamiento físicos diferentes en el sistema de almacenamiento distribuido, cada nodo de almacenamiento físico de los nodos de almacenamiento físicos diferentes incluye múltiples nodos de almacenamiento virtuales, y se despliegan nodos de almacenamiento virtuales con números de serie consecutivos en los diferentes nodos de almacenamiento físicos de acuerdo con un segundo modo de ordenamiento que es conforme con la regla preestablecida. El procesador 1210 determina, utilizando el algoritmo aleatorio, un número de serie de un primer nodo de almacenamiento virtual correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación, y determina números de serie de por lo menos M+1 nodos de almacenamiento virtuales en base al número de serie del primer nodo de almacenamiento virtual y de acuerdo con el primer modo de ordenamiento; y el procesador 1210 almacena por separado los por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en nodos de almacenamiento virtuales correspondientes a los números de serie de los por lo menos M+1 nodos de almacenamiento virtuales.

El primer modo de ordenamiento o el segundo modo de ordenamiento se refieren a un procedimiento de ordenamiento secuencial a un intervalo fijo.

Para las operaciones y funciones de cada módulo del aparato de almacenamiento distribuido 1200, se puede hacer referencia a las etapas 110, 120 y 130 del procedimiento de la figura 1. No se dan a conocer en este caso detalles adicionales para evitar su repetición.

Un experto en la materia puede estar al tanto de que, en combinación cuyos ejemplos descritos en las realizaciones dadas a conocer en esta memoria descriptiva, las unidades y etapas de algoritmos se pueden implementar mediante hardware electrónico, o mediante una combinación de software informático y hardware electrónico. Que las funciones se lleven a cabo mediante hardware o software depende de las aplicaciones particulares y de las condiciones de imposiciones de diseño de la solución técnica. Un experto en la materia puede utilizar diferentes procedimientos para implementar las funciones descritas para cada aplicación particular, pero no se deberá considerar que la implementación rebasa el alcance de la presente invención.

Un experto en la materia puede comprender fácilmente que, con el objetivo de una descripción cómoda y breve, para un proceso de trabajo detallado del sistema, del aparato y de la unidad anteriores, se puede hacer referencia a un proceso correspondiente en las realizaciones de procedimiento anteriores, y no se vuelven a describir los detalles en la presente memoria.

En las diversas realizaciones dadas a conocer en la presente solicitud, se debe entender que el sistema, el aparato y el procedimiento dados a conocer se pueden implementar de otras maneras. Por ejemplo, la realización de aparato descrita es tan sólo a modo de ejemplo. Por ejemplo, la división en unidades es una división funcional meramente lógica y puede existir otra división en una implementación real. Por ejemplo, una serie de unidades o componentes se pueden combinar o integrar en otro sistema, o algunas características pueden ser ignoradas o no ejecutadas. Además, los acoplamientos mutuos o acoplamientos directos o conexiones de comunicación representadas o discutidas se pueden implementar a través de algunas interfaces. Los acoplamientos indirectos o conexiones de comunicación entre los aparatos o unidades se pueden implementar de forma electrónica, mecánica u otras.

Las unidades descritas como partes independientes pueden o no ser físicamente independientes, y las partes mostradas como unidades pueden o no ser unidades físicas, pueden estar situadas en una posición o pueden estar distribuidas en una serie de unidades de red. Una parte o la totalidad de las unidades se pueden seleccionar en función de las necesidades reales para conseguir los objetivos de las soluciones de las realizaciones.

Además, las unidades funcionales en las realizaciones de la presente invención pueden estar integradas en una unidad de proceso, o cada una de las unidades puede existir de manera físicamente independiente, o dos o más unidades pueden estar integradas en una unidad.

Cuando las funciones están implementadas en forma de una unidad funcional de software y son vendidas o utilizadas como un producto independiente, las funciones pueden estar almacenadas en un medio de almacenamiento legible por ordenador. Comprendiéndose lo anterior, las soluciones técnicas de la presente invención esencialmente, o la parte que contribuye a la técnica anterior, o una parte de las soluciones técnicas, se pueden implementar en forma de un producto de software. El producto de software informático se almacena en un medio de almacenamiento e incluye varias instrucciones para instruir a un dispositivo informático (que puede ser un ordenador personal, un servidor o un dispositivo de red) para llevar a cabo la totalidad o parte de las etapas de los procedimientos descritos en las realizaciones de la presente invención. El anterior medio de almacenamiento incluye: cualquier medio que pueda almacenar código de programa, tal como una unidad flash USB, un disco duro extraíble, una memoria de sólo lectura en (ROM, Read-Only Memory), una memoria de acceso aleatorio (RAM, Random Access Memory), un disco magnético o un disco óptico.

Las anteriores descripciones son tan sólo realizaciones específicas de la presente invención, pero no están destinadas a limitar el alcance de protección de la presente invención. Cualquier variación o sustitución que conciba fácilmente a un experto en la materia dentro del alcance técnico dado a conocer en la presente invención caerá dentro del alcance de protección de la presente invención. Por lo tanto, el alcance de protección de la presente invención estará sujeto al alcance de protección de las reivindicaciones.

REIVINDICACIONES

1. Un procedimiento de almacenamiento distribuido, que comprende:

dividir un archivo de datos para generar K segmentos de datos, dividir cada segmento de datos de los K segmentos de datos para generar M bloques de datos para cada segmento de datos y llevar a cabo codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación (110);

determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme a una regla preestablecida (120); y

almacenar por separado por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros positivos (130).

2. El procedimiento de almacenamiento distribuido según la reivindicación 1, en el que la determinación, utilizando un algoritmo aleatorio, de un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación comprende:

llevar a cabo un cálculo de hash sobre un identificador de un bloque de datos o de un bloque de verificación de los M bloques de datos o de los N bloques de verificación para generar un valor de clave; y

determinar, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utilizar el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

3. El procedimiento de almacenamiento distribuido según la reivindicación 1, en el que la determinación, utilizando un algoritmo aleatorio, de un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación comprende:

llevar a cabo un cálculo de hash sobre un identificador del segmento de datos que está dividido en los M bloques de datos, para generar un valor de clave; y

determinar, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utilizar el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

4. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 3, en el que la determinación de por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida comprende:

determinar M+N nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento (370); y

almacenar por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos comprende:

almacenar los M bloques de datos y los N bloques de verificación en los M+N nodos de almacenamiento físicos diferentes (380).

5. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 4, en el que la determinación de por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida comprende:

mapear el primer nodo de almacenamiento físico a un bloque de los por lo menos M+1 bloques, y determinar, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a otros por lo menos M bloques de los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos comprenden el primer nodo de almacenamiento físico;

o

determinar, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos no comprenden el primer nodo de almacenamiento físico.

6. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 5, en el que los diferentes nodos de almacenamiento físicos son discos duros, y el número de discos duros es mayor o igual que M+1;

o

los diferentes nodos de almacenamiento físicos son servidores, donde el número de servidores es mayor o igual que M+1;

o

los diferentes nodos de almacenamiento físicos son bastidores, donde el número de bastidores es mayor o igual que M+1;

7. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 6, que comprende además:

en caso de que uno de los por lo menos M+1 diferentes nodos de almacenamiento físicos que almacena por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación falle, restablecer datos del nodo de almacenamiento físico que falla de los por lo menos M+1 diferentes nodos de almacenamiento físicos en un nodo de almacenamiento de respaldo en caliente en un sistema de almacenamiento distribuido en el que está situado el nodo de almacenamiento físico.

8. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 6, que comprende además:

en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle y si L no es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es opuesto al sentido de una secuencia en el primer modo de ordenamiento, y si L es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es el mismo que el sentido de la secuencia en el primer modo de ordenamiento;

o

en caso de que falle un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido.

o

en caso de que falle un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido.

9. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 8, que comprende además:

en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos y si P no es mayor que (M+N)/2, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es el mismo que el sentido de una secuencia en el primer modo de ordenamiento; y si P es mayor que (M+N)/2, migrar un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de

ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es opuesto al sentido de la secuencia en el primer modo de ordenamiento, y P es un entero;

o

en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido;

o

en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico en el que está situado un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido.

10. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 9, que comprende además:

cuando se tiene que leer el archivo de datos, determinar, utilizando el algoritmo aleatorio, el primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento;

leer por lo menos M bloques desde los por lo menos M nodos de almacenamiento físicos diferentes, donde los M bloques leídos comprenden los M bloques de datos o comprenden algunos bloques de datos de los M bloques de datos y algunos bloques de verificación de los N bloques de verificación; y

llevar a cabo descodificación y verificación inversa sobre los por lo menos M bloques para generar M bloques de datos descodificados, y combinar los M bloques de datos descodificados para generar el archivo de datos.

11. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 9, en el que el archivo de datos se adquiere desde fuera del sistema de almacenamiento distribuido, o el archivo de datos se adquiere a partir de algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido.

12. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 11, en el que el algoritmo de redundancia es un algoritmo de codificación de borrado, y el algoritmo aleatorio es un algoritmo de tabla de hash distribuida.

13. El procedimiento de almacenamiento distribuido según cualquiera de las reivindicaciones 1 a 12, en el que los nodos de almacenamiento físicos diferentes son nodos de almacenamiento físicos diferentes del sistema de almacenamiento distribuido, cada nodo de almacenamiento físico de los nodos de almacenamiento físicos diferentes comprende múltiples nodos de almacenamiento virtuales, y se despliegan nodos de almacenamiento virtuales con números de serie consecutivos en los nodos de almacenamiento físicos diferentes de acuerdo con un segundo modo de ordenamiento que es conforme con una regla preestablecida;

determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme a una regla preestablecida comprende:

determinar, utilizando el algoritmo aleatorio, el número de serie de un primer nodo de almacenamiento virtual correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación, y determinar números de serie de por lo menos M+1 nodos de almacenamiento virtuales en base al número de serie del primer nodo de almacenamiento virtual y de acuerdo con el primer modo de ordenamiento; y

almacenar por separado por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos comprende:

almacenar los por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en nodos de almacenamiento virtuales correspondientes a los números de serie de los por lo menos M+1 nodos de almacenamiento virtuales.

14. Un aparato de almacenamiento distribuido, que comprende:

un módulo de generación (710; 810), configurado para dividir un archivo de datos para generar K segmentos de datos, dividir cada segmento de datos de los K segmentos de datos para generar M bloques de datos para cada segmento de datos, y llevar a cabo codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación;

un módulo de determinación (720, 820) configurado para determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento que es conforme con una regla preestablecida; y

un módulo de almacenamiento (730; 830) configurado para almacenar por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros positivos.

15. El aparato de almacenamiento distribuido según la reivindicación 14, en el que el módulo de determinación (720; 820) lleva a cabo un cálculo de hash sobre el identificador de un bloque de datos o un bloque de verificación de los M bloques de datos o de los N bloques de verificación para generar un valor de clave, determina, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utiliza el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

16. El aparato de almacenamiento distribuido según la reivindicación 14, en el que el módulo de determinación (720; 820) está configurado para llevar a cabo un cálculo de hash sobre un identificador del segmento de datos que está dividido en los M bloques de datos, con el fin de generar un valor de clave, determinar, de acuerdo con el valor de clave, un nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y utilizar el nodo de almacenamiento físico como el primer nodo de almacenamiento físico.

17. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 16, en el que el módulo de determinación (720; 820) determina M+N nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, y el módulo de almacenamiento (730; 830) almacena los M bloques de datos y los N bloques de verificación en los M+N nodos de almacenamiento físicos diferentes.

18. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 17, en el que el módulo de determinación (720; 820) mapea el primer nodo de almacenamiento físico a un bloque de los por lo menos M+1 bloques, y determina, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a otros por lo menos M bloques de los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos comprenden el primer nodo de almacenamiento físico; o determina, en base a la posición del primer nodo de almacenamiento físico y de acuerdo con el primer modo de ordenamiento, nodos de almacenamiento físicos correspondientes a los por lo menos M+1 bloques, donde los M+1 diferentes nodos de almacenamiento físicos no comprenden el primer nodo de almacenamiento físico.

19. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 18, en el que los diferentes nodos de almacenamiento físicos son discos duros, y el número de discos duros es mayor o igual que M+1; o los diferentes nodos de almacenamiento físicos son servidores, donde el número de servidores es mayor o igual que M+1; o los diferentes nodos de almacenamiento físicos son bastidores, donde el número de bastidores es mayor o igual que M+1.

20. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 19, que comprende además:

un módulo de restablecimiento (850), configurado para, en caso de que uno de los por lo menos M+1 diferentes nodos de almacenamiento físicos que almacenan por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación falle, restablecer datos del nodo de almacenamiento físico que falla de los por lo menos M+1 diferentes nodos de almacenamiento físicos en un nodo de almacenamiento de respaldo en caliente en un sistema de almacenamiento distribuido en el que está situado el nodo de almacenamiento físico.

21. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 19, que comprende además:

un módulo de restablecimiento (850), configurado para, en caso de que un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos falle y si L no es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es opuesto al sentido de una secuencia en el primer modo de ordenamiento, y si L es mayor que (M+N)/2, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es el mismo que el sentido de la secuencia en el primer modo de ordenamiento;

o

un módulo de restablecimiento (850), configurado para, en caso de que falle un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques anteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un primer sentido.

o

un módulo de restablecimiento (850), configurado para, en caso de que falle un nodo de almacenamiento físico de un bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar secuencialmente el bloque L-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y los bloques posteriores al bloque L-ésimo a nodos de almacenamiento físicos adyacentes en un segundo sentido.

22. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 21, que comprende además:

un módulo de expansión de la capacidad (860), configurado para, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos y si P no es mayor que (M+N)/2, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido, donde el primer sentido es el mismo que el sentido de una secuencia en el primer modo de ordenamiento; y si P es mayor que (M+N)/2, migrar un bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido, donde el segundo sentido es opuesto al sentido de la secuencia en el primer modo de ordenamiento, y P es un entero;

o

un módulo de expansión de la capacidad (860), configurado para, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico de un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques anteriores al bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un primer sentido;

o

un módulo de expansión de la capacidad (860), configurado para, en caso de que se añada un nuevo nodo de almacenamiento físico después de un nodo de almacenamiento físico en el que está situado un bloque P-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento y almacenados en los por lo menos M+1 diferentes nodos de almacenamiento físicos, migrar el bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento al nuevo nodo de almacenamiento físico, y migrar secuencialmente bloques

posteriores al bloque (P+1)-ésimo de los M bloques de datos y de los N bloques de verificación que están ordenados de acuerdo con el primer modo de ordenamiento a nodos de almacenamiento físicos adyacentes en un segundo sentido.

23. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 22, en el que el módulo de determinación está configurado además para, cuando se tiene que leer el archivo de datos, determinar, utilizando el algoritmo aleatorio, el primer nodo de almacenamiento físico correspondiente a un bloque de datos o a un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M nodos de almacenamiento físicos diferentes en base al primer nodo de almacenamiento físico determinado y de acuerdo con el primer modo de ordenamiento, y el aparato de almacenamiento distribuido comprende además:

un módulo de lectura, configurado para leer por lo menos M bloques a partir de los M diferentes nodos de almacenamiento físicos, donde los M bloques leídos comprenden los M bloques de datos o comprenden algunos bloques de datos de los M bloques de datos y algunos bloques de verificación de los N bloques de verificación, donde el módulo de generación está configurado además para llevar a cabo descodificación y verificación inversa sobre los por lo menos M bloques para generar M bloques de datos descodificados, y combinar los M bloques de datos descodificados para generar el archivo de datos.

24. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 22, que comprende además:

un módulo de adquisición, configurado para adquirir el archivo de datos desde fuera del sistema de almacenamiento distribuido, o adquirir el archivo de datos desde algunos nodos de almacenamiento físicos del sistema de almacenamiento distribuido.

25. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 24, en el que el algoritmo de redundancia es un algoritmo de codificación de borrado, y el algoritmo aleatorio es un algoritmo de tabla de hash distribuida.

26. El aparato de almacenamiento distribuido según cualquiera de las reivindicaciones 14 a 25, en el que los nodos de almacenamiento físicos diferentes son nodos de almacenamiento físicos diferentes en el sistema de almacenamiento distribuido, cada nodo de almacenamiento físico de los nodos de almacenamiento físicos diferentes comprende múltiples nodos de almacenamiento virtuales, y los nodos de almacenamiento virtuales con números de serie consecutivos se despliegan en los nodos de almacenamiento físicos diferentes de acuerdo con un segundo modo de ordenamiento que es conforme con una regla preestablecida; el módulo de determinación (720; 820) determina, utilizando el algoritmo aleatorio, un número de serie de un primer nodo de almacenamiento virtual correspondiente a un bloque de los M bloques de datos y de los N bloques de verificación, y determina números de serie de por lo menos M+1 nodos de almacenamiento virtuales en base al número de serie del primer nodo de almacenamiento virtual y de acuerdo con el primer modo de ordenamiento; y el módulo de almacenamiento (730; 830) almacena los por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en nodos de almacenamiento virtuales correspondientes a los números de serie de los por lo menos M+1 nodos de almacenamiento virtuales.

27. Un sistema de almacenamiento distribuido, caracterizado por que el sistema comprende:

un cliente (910);

múltiples nodos de almacenamiento físicos (920); y

el aparato de almacenamiento distribuido (930) según las reivindicaciones 14 a 26, en el que el aparato de almacenamiento distribuido lee/almacena un archivo de datos de un usuario en/desde los múltiples nodos de almacenamiento físicos de acuerdo con una solicitud de almacenamiento/lectura enviada por el usuario a través del cliente.

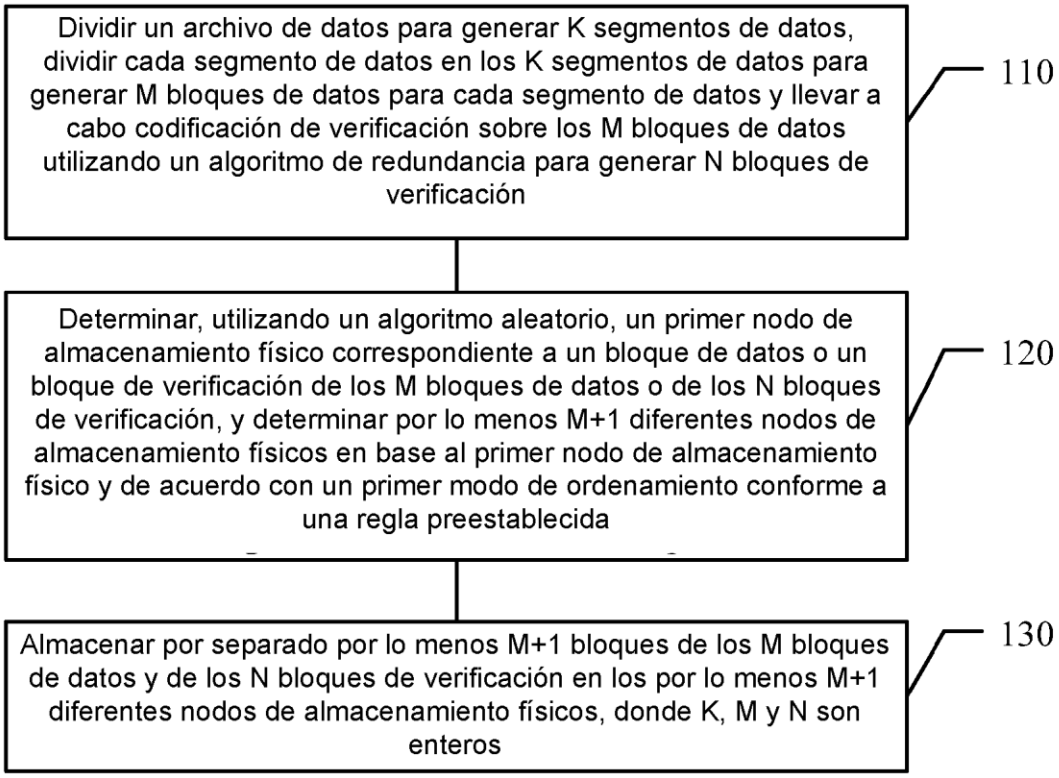
Dividir un archivo de datos para generar K segmentos de datos, dividir cada segmento de datos en los K segmentos de datos para generar M bloques de datos para cada segmento de datos y llevar a cabo codificación de verificación sobre los M bloques de datos utilizando un algoritmo de redundancia para generar N bloques de verificación
110
Determinar, utilizando un algoritmo aleatorio, un primer nodo de almacenamiento físico correspondiente a un bloque de datos o un bloque de verificación de los M bloques de datos o de los N bloques de verificación, y determinar por lo menos M+1 diferentes nodos de almacenamiento físicos en base al primer nodo de almacenamiento físico y de acuerdo con un primer modo de ordenamiento conforme a una regla preestablecida
120
Almacenar por separado por lo menos M+1 bloques de los M bloques de datos y de los N bloques de verificación en los por lo menos M+1 diferentes nodos de almacenamiento físicos, donde K, M y N son enteros
130

FIG. 1

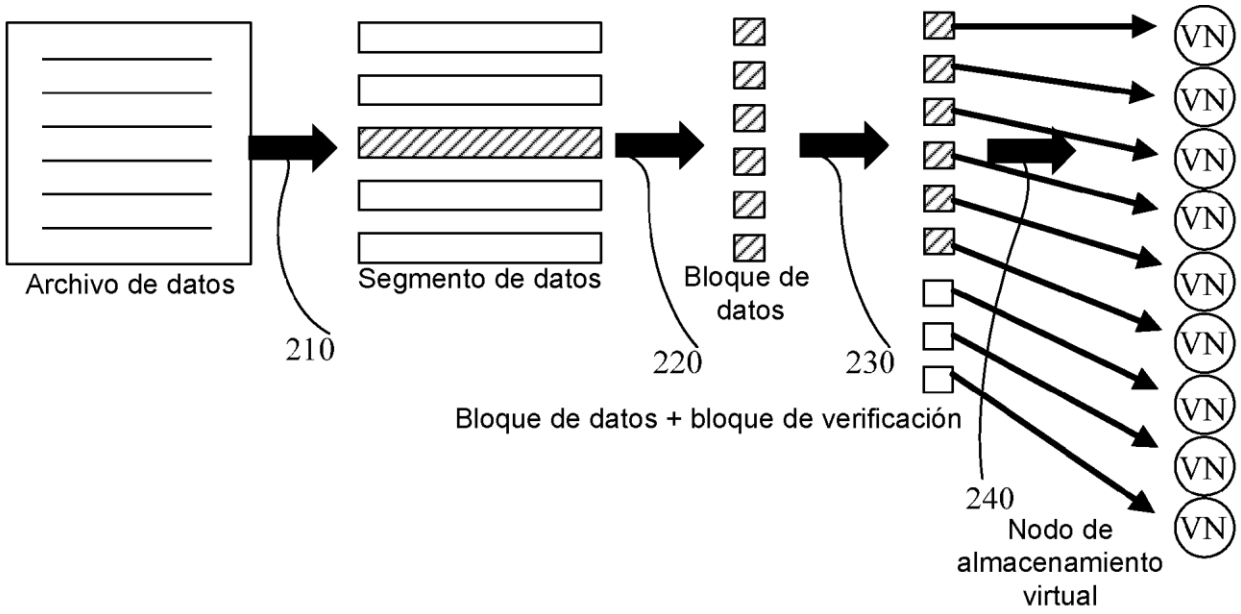
Archivo de datos
210
Segmento de datos
220
Bloque de datos
230
Bloque de datos + bloque de verificación
VN
VN
VN
VN
VN
VN
VN
VN
VN
240
Nodo de almacenamiento virtual

FIG. 2

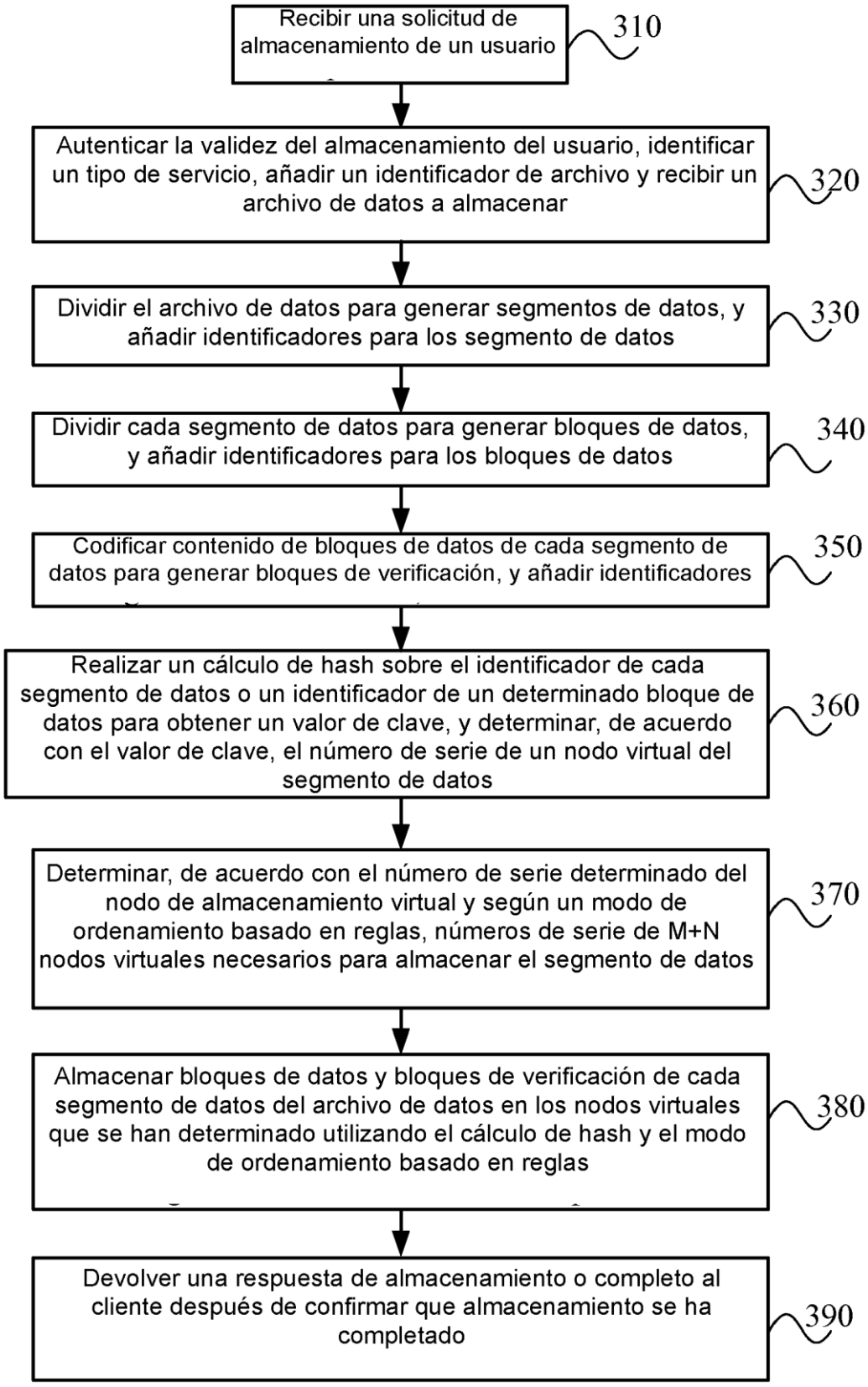
Recibir una solicitud de almacenamiento de un usuario
310
Autenticar la validez del almacenamiento del usuario, identificar un tipo de servicio, añadir un identificador de archivo y recibir un archivo de datos a almacenar
320
Dividir el archivo de datos para generar segmentos de datos, y añadir identificadores para los segmento de datos
330
Dividir cada segmento de datos para generar bloques de datos, y añadir identificadores para los bloques de datos
340
Codificar contenido de bloques de datos de cada segmento de datos para generar bloques de verificación, y añadir identificadores
350
Realizar un cálculo de hash sobre el identificador de cada segmento de datos o un identificador de un determinado bloque de datos para obtener un valor de clave, y determinar, de acuerdo con el valor de clave, el número de serie de un nodo virtual del segmento de datos
360
Determinar, de acuerdo con el número de serie determinado del nodo de almacenamiento virtual y según un modo de ordenamiento basado en reglas, números de serie de M+N nodos virtuales necesarios para almacenar el segmento de datos
370
Almacenar bloques de datos y bloques de verificación de cada segmento de datos del archivo de datos en los nodos virtuales que se han determinado utilizando el cálculo de hash y el modo de ordenamiento basado en reglas
380
Devolver una respuesta de almacenamiento o completo al cliente después de confirmar que almacenamiento se ha completado
390

FIG. 3

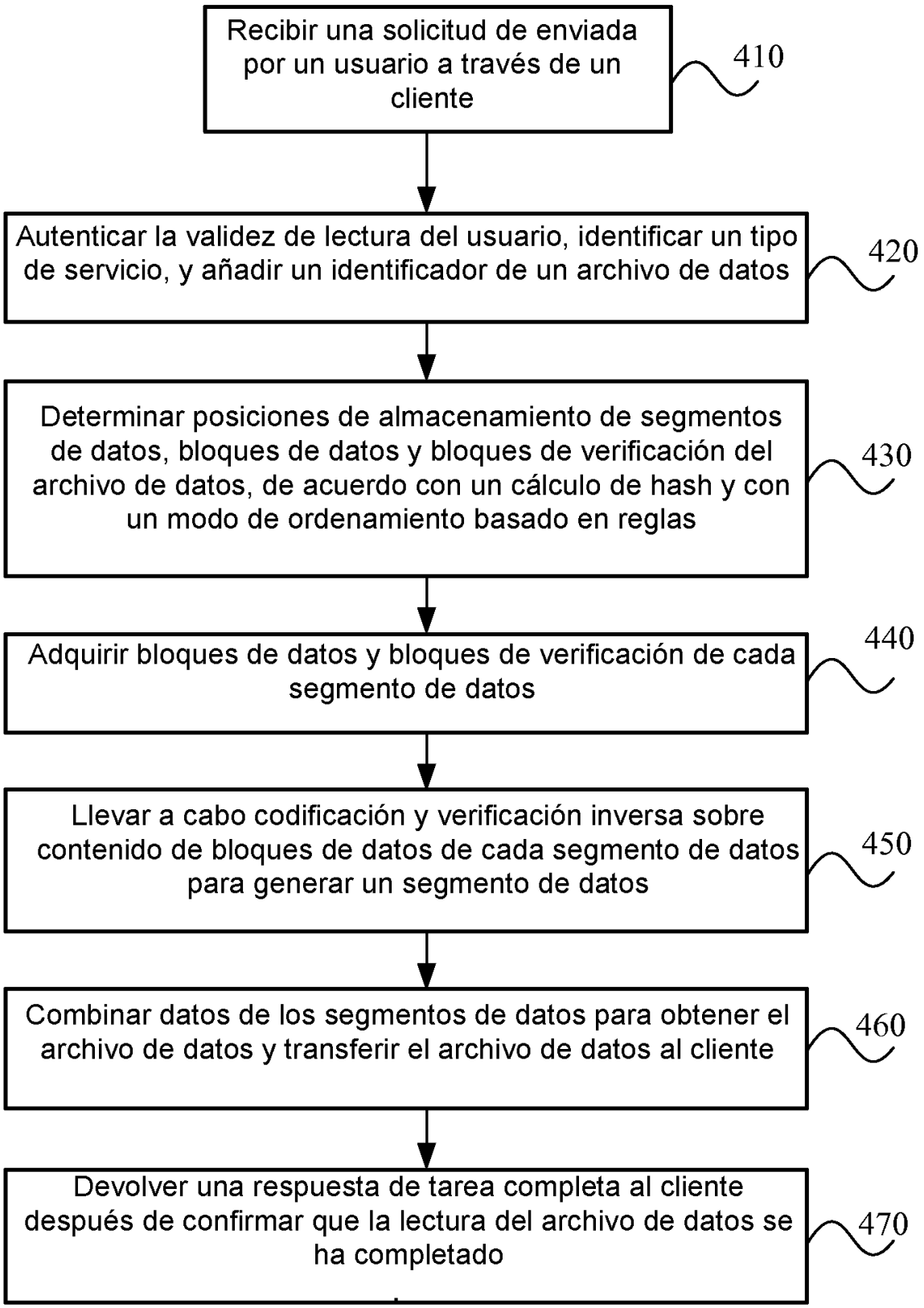
Recibir una solicitud de enviada por un usuario a través de un cliente
410
Autenticar la validez de lectura del usuario, identificar un tipo de servicio, y añadir un identificador de un archivo de datos
420
Determinar posiciones de almacenamiento de segmentos de datos, bloques de datos y bloques de verificación del archivo de datos, de acuerdo con un cálculo de hash y con un modo de ordenamiento basado en reglas
430
Adquirir bloques de datos y bloques de verificación de cada segmento de datos
440
Llevar a cabo codificación y verificación inversa sobre contenido de bloques de datos de cada segmento de datos para generar un segmento de datos
450
Combinar datos de los segmentos de datos para obtener el archivo de datos y transferir el archivo de datos al cliente
460
Devolver una respuesta de tarea completa al cliente después de confirmar que la lectura del archivo de datos se ha completado
470

FIG. 4

									Disco duro i-1	Disco duro i	Disco duro i+1							
Segmento 1									1	2	3	4	5	6	7	8	9	
Segmento 2								1	2	3	4	5	6	7	8	9		
Segmento 3							1	2	3	4	5	6	7	8	9			
Segmento 4						1	2	3	4	5	6	7	8	9				
Segmento 5					1	2	3	4	5	6	7	8	9					
Segmento 6				1	2	3	4	5	6	7	8	9						
Segmento 7			1	2	3	4	5	6	7	8	9							
Segmento 8		1	2	3	4	5	6	7	8	9								
Segmento 9	1	2	3	4	5	6	7	8	9									

FIG. 5A

										Disco duro i-1	Disco duro i	Disco duro i'	Disco duro i+1						
Segmento 1										1	2	3	4	5	6	7	8	9	
Segmento 2								1		2	3	4	5	6	7	8	9		
Segmento 3							1	2		3	4	5	6	7	8	9			
Segmento 4						1	2	3		4	5	6	7	8	9				
Segmento 5					1	2	3	4		5	6	7	8	9					
Segmento 6				1	2	3	4	5		6	7	8	9						
Segmento 7			1	2	3	4	5	6		7	8	9							
Segmento 8		1	2	3	4	5	6	7		8	9								
Segmento 9	1	2	3	4	5	6	7	8		9									

FIG. 5B

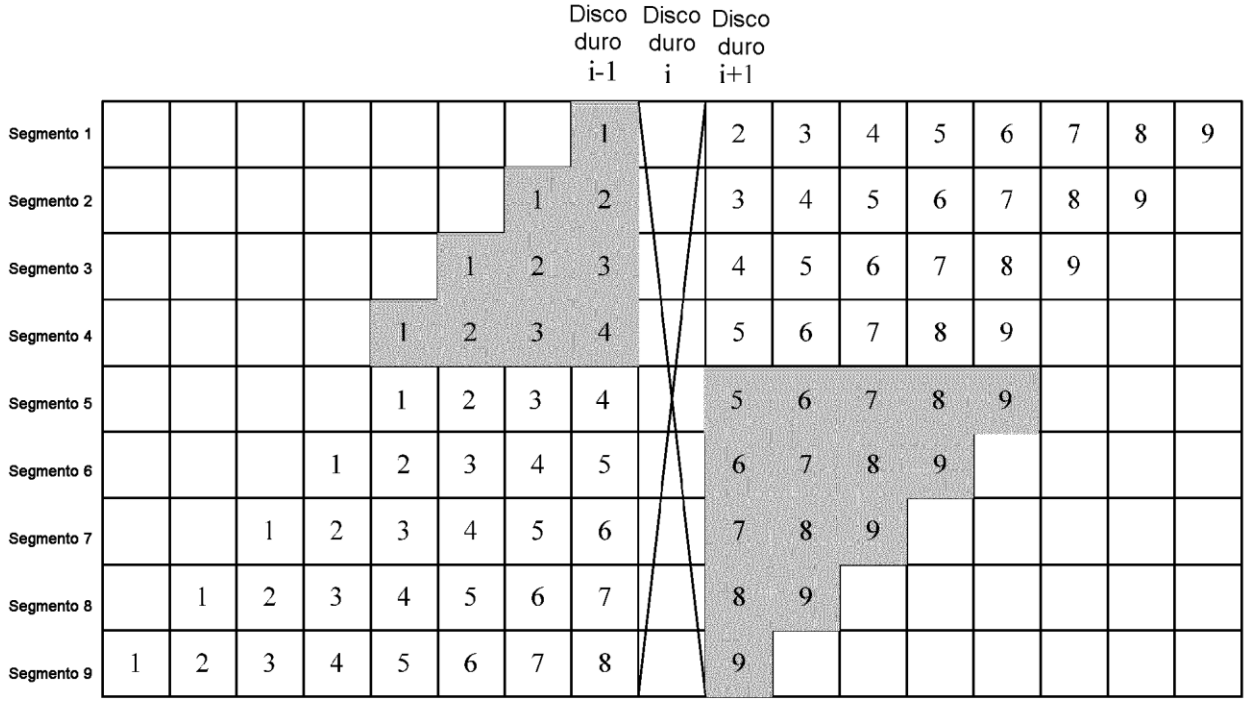
Disco duro i-1
Disco duro i
Disco duro i+1
Segmento 1
Segmento 2
Segmento 3
Segmento 4
Segmento 5
Segmento 6
Segmento 7
Segmento 8
Segmento 9

FIG. 5C

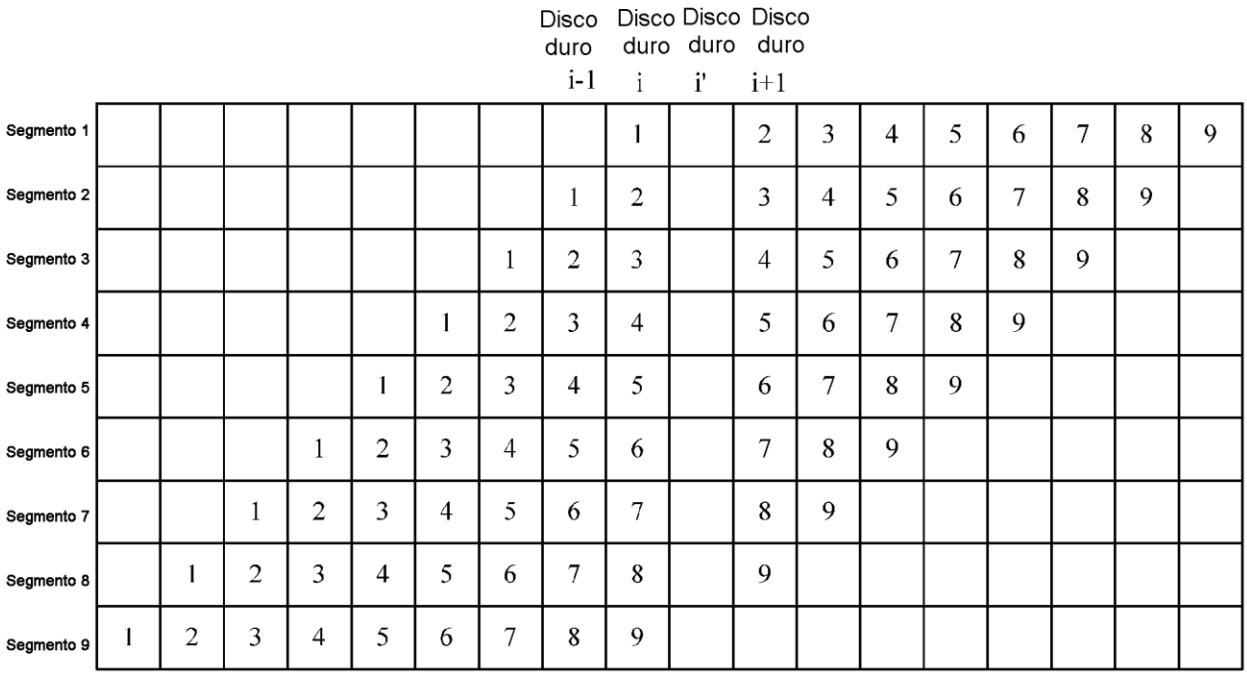
Disco duro i-1
Disco duro i
Disco duro i'
Disco duro i+1
Segmento 1
Segmento 2
Segmento 3
Segmento 4
Segmento 5
Segmento 6
Segmento 7
Segmento 8
Segmento 9

FIG. 6A

	Disco duro i-1	Disco duro i	Disco duro i'	Disco duro i+1													
Segmento 1										1	2	3	4	5	6	7	8	9
Segmento 2									1	2	3	4	5	6	7	8	9	
Segmento 3								1	2	3	4	5	6	7	8	9		
Segmento 4							1	2	3	4	5	6	7	8	9			
Segmento 5					1	2	3	4	5	6	7	8	9					
Segmento 6				1	2	3	4	5	6	7	8	9						
Segmento 7			1	2	3	4	5	6	7	8	9							
Segmento 8		1	2	3	4	5	6	7	8	9								
Segmento 9	1	2	3	4	5	6	7	8	9									

FIG. 6B

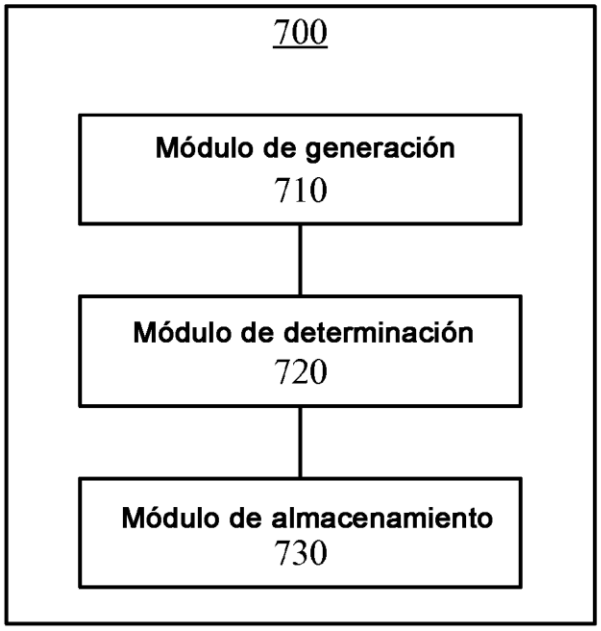

FIG. 7

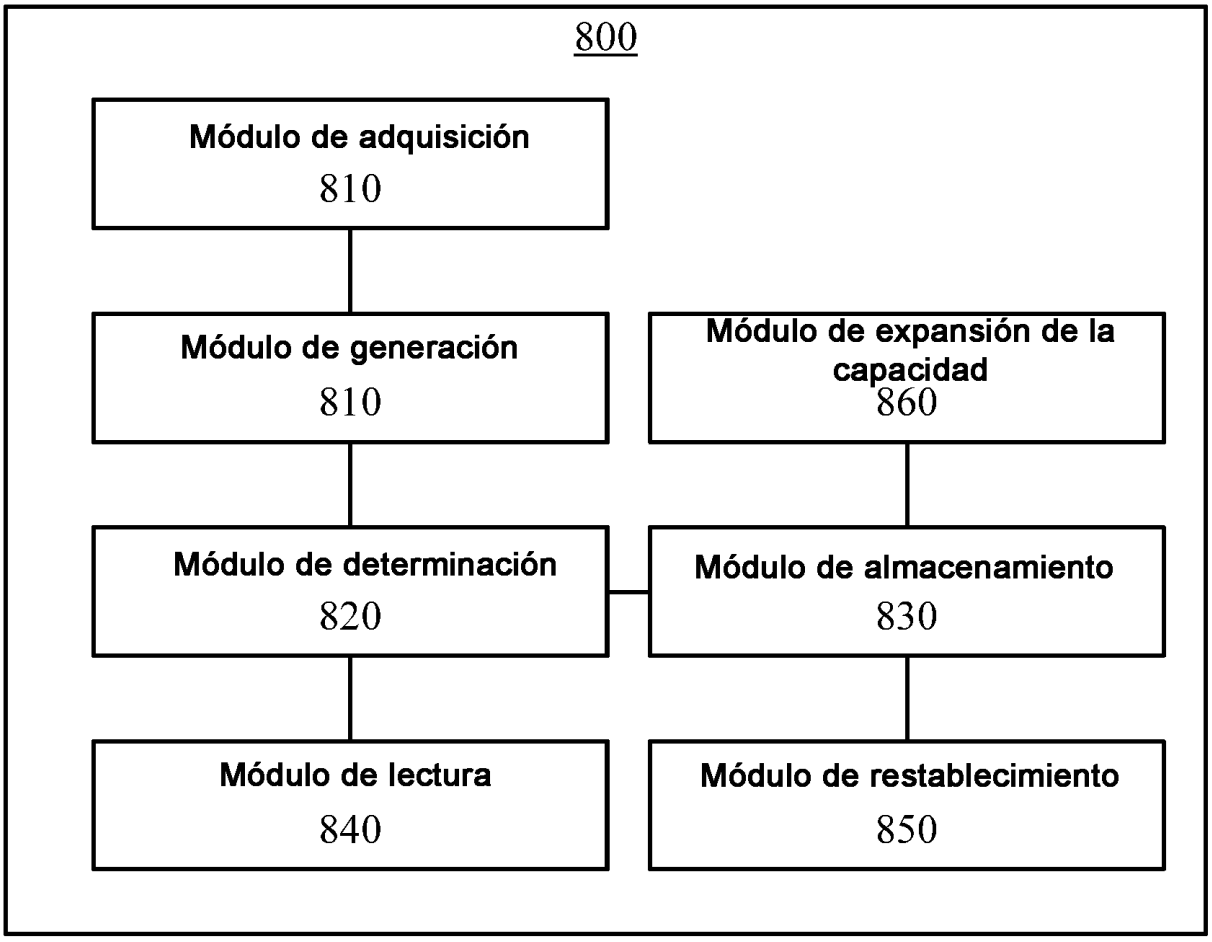
800
Módulo de adquisición
810
Módulo de generación
810
Módulo de expansión de la capacidad
860
Módulo de determinación
820
Módulo de almacenamiento
830
Módulo de lectura
840
Módulo de restablecimiento
850

FIG. 8

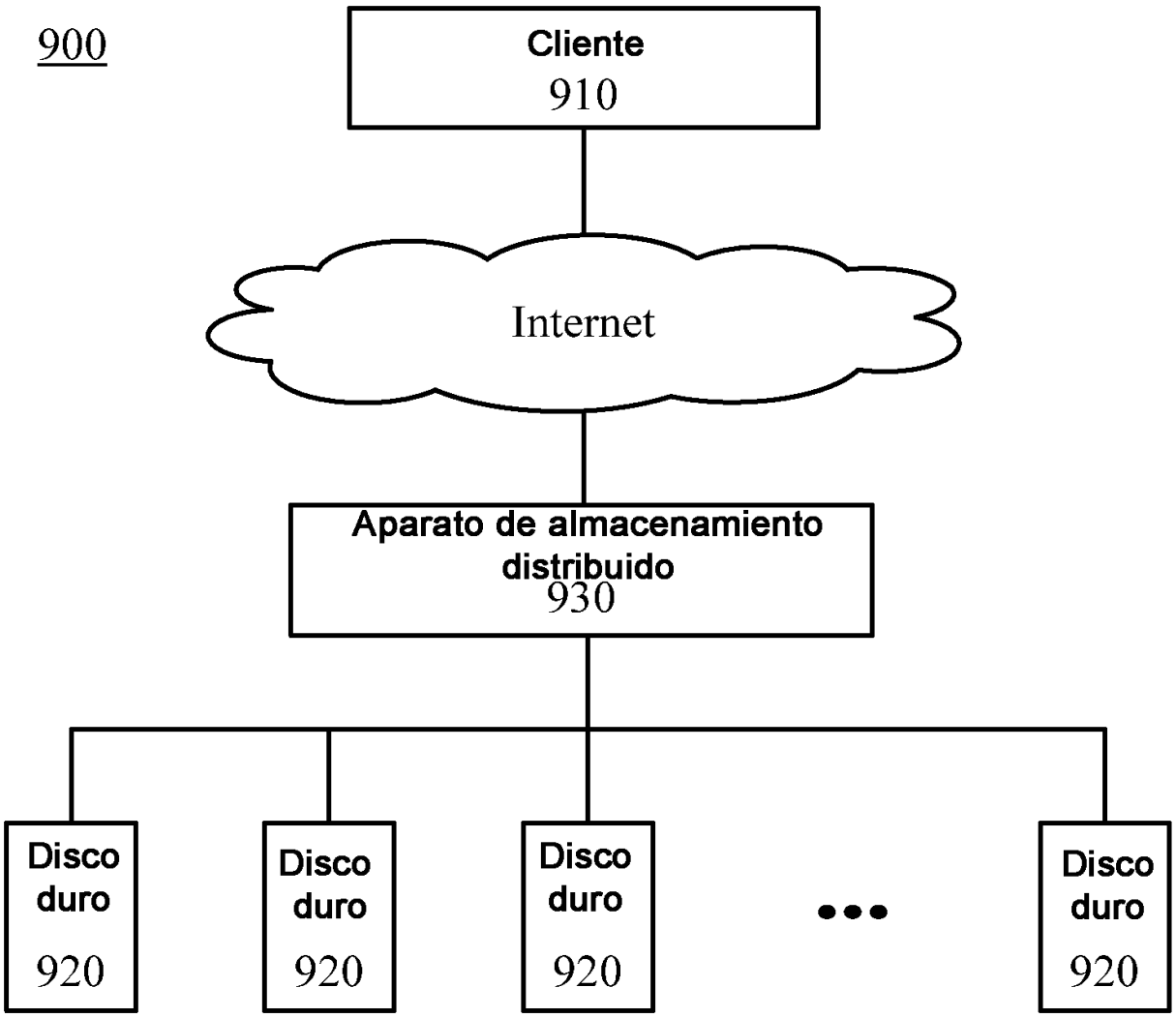
900
Cliente
910
Internet
Aparato de almacenamiento distribuido
930
Disco duro
920
Disco duro
920
Disco duro
920
Disco duro
920

FIG. 9

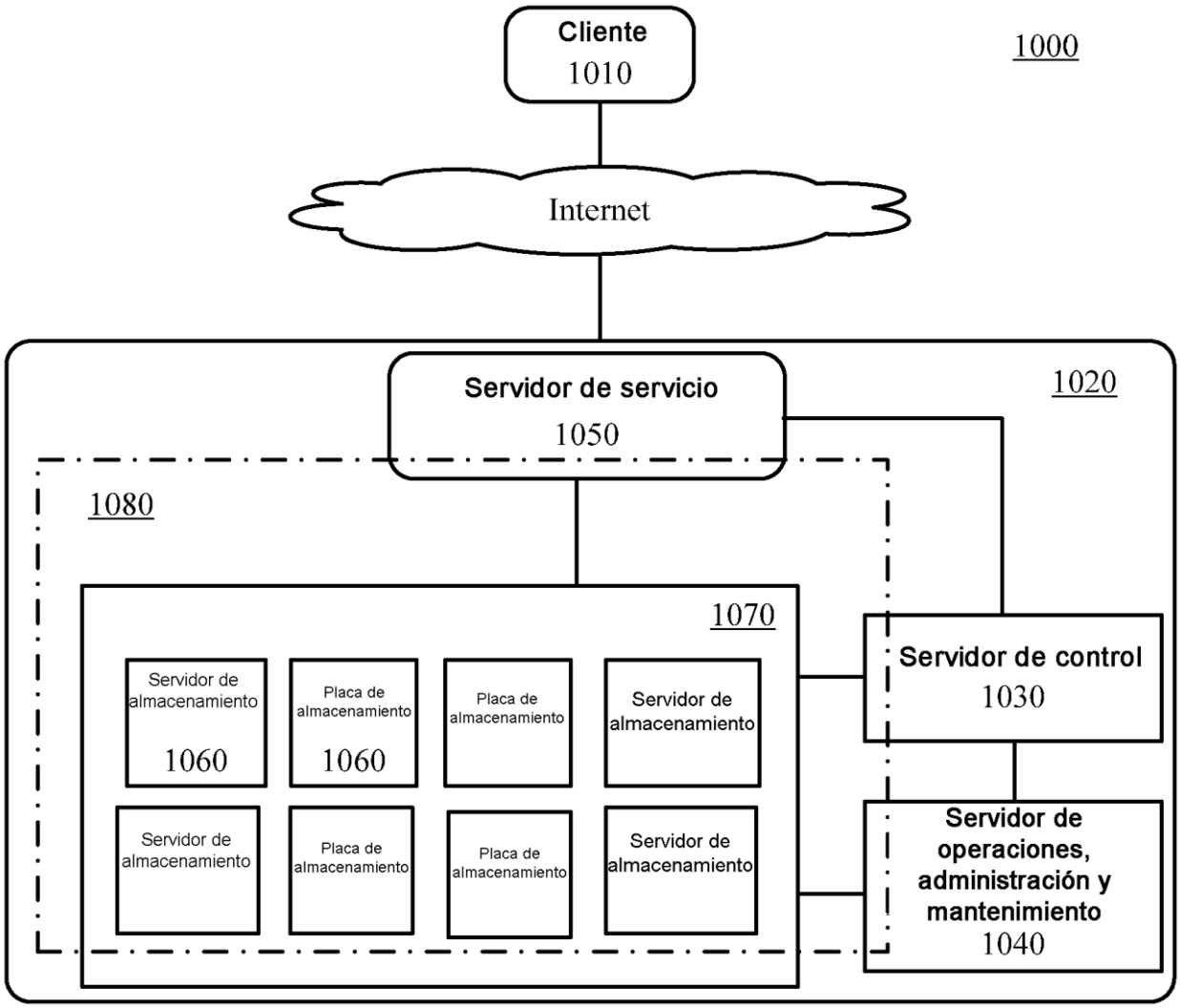

FIG. 10

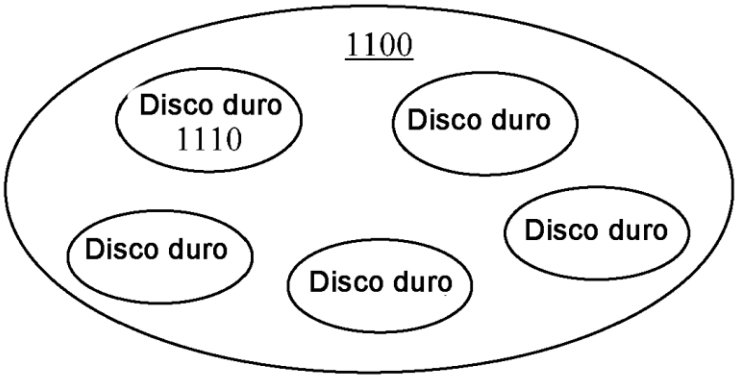

FIG. 11A

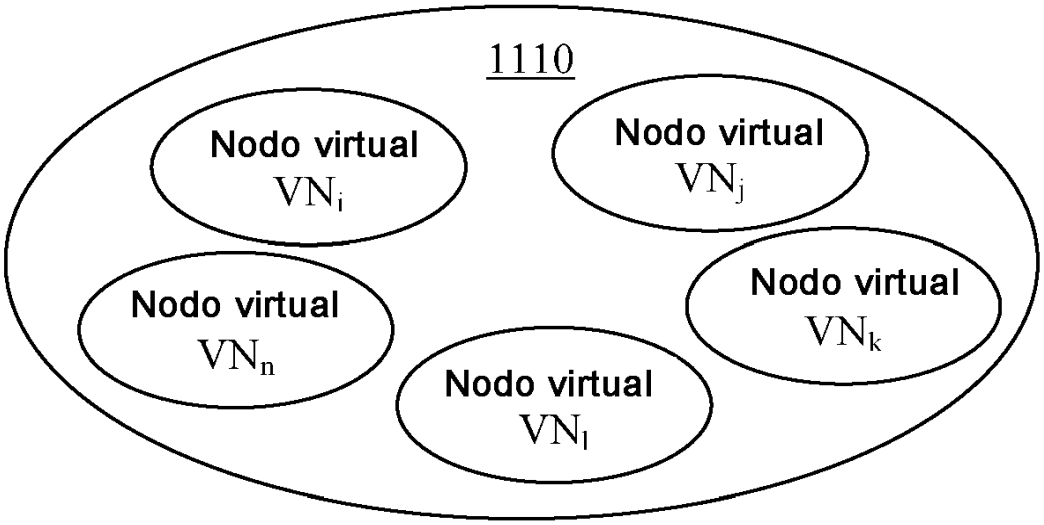

FIG. 11B

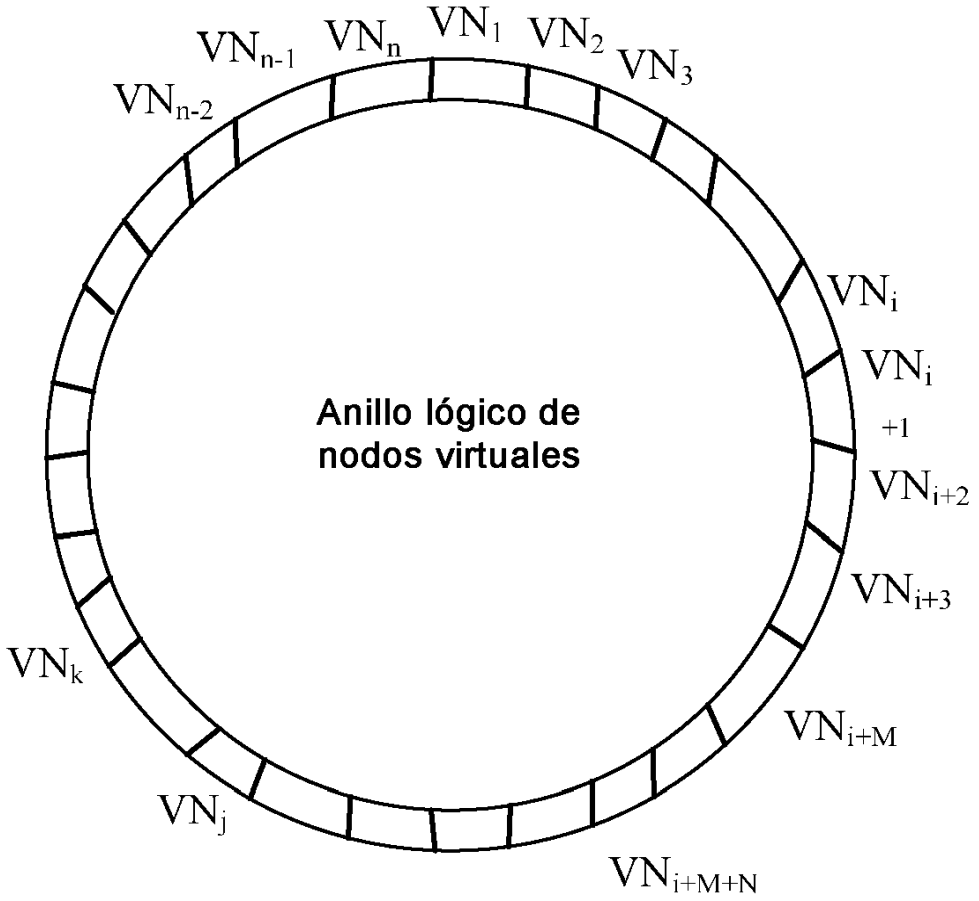

FIG. 11C

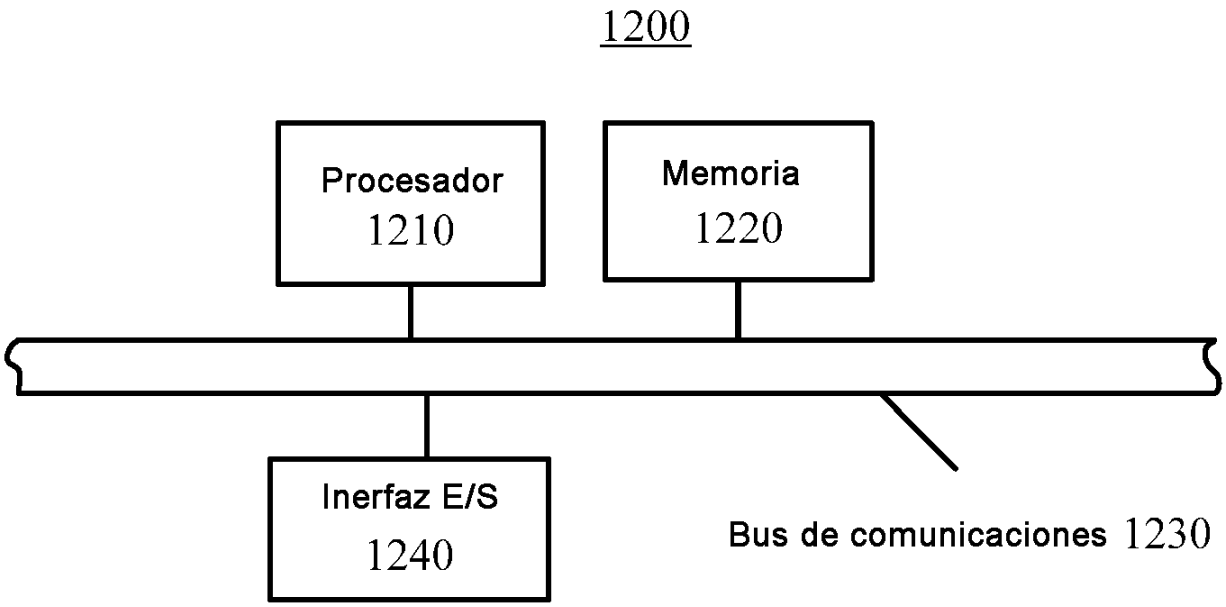
1200
Procesador
1210
Memoria
1220
Inerfaz E/S
1240
Bus de comunicaciones 1230

FIG. 12