CLAIMS

1. A data processing method, wherein the method is applied to a distributed storage system, the distributed storage system comprises a plurality of storage devices, a first storage device in the plurality of storage devices comprises a first processor, a first hard disk, and a first artificial intelligence, AI, apparatus, the first AI apparatus communicates with the first processor through a high-speed interconnect network, wherein the first storage device comprises a memory and the first AI apparatus comprises an AI processor and an AI memory and the method comprises:

receiving, by the first processor, service data, and storing the service data in the first hard disk; and

sending, by the first AI apparatus, a data obtaining request to the first processor to obtain the service data, and performing AI computing on the service data;

wherein the data obtaining request comprises a third data obtaining request, and the obtaining, by the first processor, the service data comprises:

sending, by the first processor, metadata of the service data to the first AI apparatus in response to the third data obtaining request, wherein the metadata is used to indicate an address of the service data; and

when the metadata indicates that the service data is located in a second storage device in the plurality of storage devices, sending, by the first AI apparatus, a second data access request to the second storage device, wherein the second data access request is a remote direct memory access, RDMA, request, and wherein the second data access request comprises the metadata; and receiving the service data that is sent by the second storage device in response to the second data access request; the method further comprising:

when an available capacity of the AI memory reaches a preset threshold, sending, by the AI processor, a memory application request to the first processor, wherein the available capacity of the AI memory is determined by a specified batch size, and the memory application request is used to request the first processor to obtain a segment of memory space form the memory through division and reserve the segment of memory space for the AI apparatus.

2. The method according to claim 1, wherein the method further comprises:

transmitting, by the second storage device, the service data to the first storage device through

a second network; and

transmitting an AI parameter to the first storage device through a first network, wherein the AI parameter is used to perform AI computing on the service data.

3. The method according to claim 1 or claim 2, wherein the method further comprises:

transmitting, by a second storage device in the plurality of storage devices, other service data to the first storage device through the first network; and transmitting another AI parameter through the second network, wherein the another AI parameter is used to perform AI computing on the other service data.

4. The method according to claim 1, wherein the distributed storage system further comprises a management apparatus, and the method further comprises:

receiving, by the management apparatus, a first job request; determining distribution of a to-be-trained dataset based on the first job request, wherein the dataset comprises the service data; and when determining that the service data is distributed on the first storage device, sending a first computing request to the first storage device, wherein the first computing request is used to request the first AI apparatus to perform AI computing on the service data.

5. A distributed storage system, comprising a plurality of storage devices, a first storage device in the plurality of storage devices comprises a first processor, a first hard disk, and a first artificial intelligence, AI, apparatus, the first AI apparatus communicates with the first processor through a high-speed interconnect network, wherein the first storage device comprises a memory and the first AI apparatus comprises an AI processor and an AI memory,

wherein the first AI apparatus is configured to send a third data obtaining request to the first processor to obtain service data,

wherein the first processor is configured to send metadata of service data to the first AI apparatus in response to the third data obtaining request, wherein the metadata is used to indicate an address of the service data; and

the first AI apparatus is configured to:

send a second data access request to the second storage device when the metadata indicates that the service data is located in a second storage device in the plurality of storage devices,

wherein the second data access request is a remote direct memory access, RDMA, request, and wherein the second data access request comprises the metadata;

receive the service data that is sent by the second storage device in response to the second data access request;

and perform AI computing on the service data;

wherein the AI processor is configured to: when an available capacity of the AI memory reaches a preset threshold, send a memory application request to the first processor, wherein the available capacity of the AI memory is determined by a specified batch size, and the memory application request is used to request the first processor to obtain a segment of memory space from the memory through division and reserve the segment of memory space for the AI apparatus.