DATA BACKUP METHOD AND SYSTEM, AND RELATED DEVICE

TECHNICAL FIELD

[0001] This application relates to the field of big data technologies, and in particular, to a data backup method and system, and a related device.

BACKGROUND

[0002] With development of big data technologies, more users (such as enterprises) migrate service data to a big data platform for storage. Accordingly, the user attaches more importance to a disaster recovery capability of the big data platform. In other words, the big data platform is expected to ensure that the service data stored on the big data platform is not lost when a disaster such as a device fault occurs.

[0003] Currently, the big data platform implements disaster recovery for the service data of the user through a data backup system. The data backup system includes a primary cluster and a secondary cluster. The primary cluster may process the service data of the user by using a component, for example, encapsulate and store the service data of the user by using the component. Different components may process different types of service data of a same user. For example, the primary cluster may store service data such as audio, video, and image of the user by using a component 1, and store service data of the user in a form of a table by using a component 2. Generally, the primary cluster may periodically back up service data processed by each component to a secondary site, so that a secondary cluster may continue to provide a business service for the user based on the backed up service data after the primary cluster is faulty. Therefore, after a disaster recovery switchover, how to avoid, as much as possible, that the service data backed up to the secondary cluster affects quality of the provided business service becomes an urgent problem to be resolved currently.

[0003a] US 9,275,060 B1 discloses a data protection agent or server running on a computing device receives a cluster configuration of a high availability cluster. The data protection agent or server identifies highly available data of an application running on the high availability cluster based on the clustering. The data protection agent or server then implements a data protection policy that backs up the highly available data.

[0003b] US 10,089,187 B1 discloses implementations for scalable cloud backup. A coordinator

process can manage worker processes on nodes to package file system data that is targeted for cloud backup into node local upload objects. File data can be arranged into distinct block offsets of the node local upload object. A set of metadata tables can be generated that characterize each file that is backed up as well as file block location information for each data block. The node local upload objects can be uploaded to a cloud service provider. The set of metadata tables generated by the worker process can be coalesced into a global set of metadata tables that describe the data that has been backed up. In one implementation, after an initial cloud backup has occurred, a snapshot service of the file system can be used to incrementally backup blocks of the file that have been changed.

[0003c] CN 112527567 A discloses a system disaster tolerance method, device and equipment and a storage medium, and relates to the technical field of cloud computing and applets. The specific implementation scheme is as follows: determining whether a backup cluster in the system is available or not in response to the determination of a main cluster fault in the system; in response to determining that the backup cluster is available, modifying cluster configuration information; and outputting the modified cluster configuration information to a user of the main cluster, so that the user uses the backup cluster according to the modified cluster configuration information. According to the implementation mode, the high availability of the system can be improved.

[0003d] EP 1 921 540 A2 discloses: The reconfiguration controller selects for each data whether or not to execute remote copying of data from the first storage device to the second storage device. Each data sent from the host to be stored in the first storage device contains a copy policy tag that defines the policy relating to remote copying. The reconfiguration controller determines whether or not remote copying of the each data is necessary based on the copy policy tag of the each data, and selectively transmits the data requiring remote copying to the second storage device.

[0003e] Anonymous: "Synchronize the Cluster Node Time", 24 August 2020 (2020-08-24), pages 1-2, Retrieved from the Internet: URL:https://techdocs.broadcom.com/us/en/ca-enterprise-software/intelligent-automation/automic-process-automation/4-3/installing/install-and-configure/orchestrators/synchronize-the-cluster-node-time.html discloses methods for synchronizing the time of all nodes in a cluster.

[0003f] US 2021/011813 A1 discloses an information backup method and a related device, to ensure continuity of a user service. The method is applied to a communications system including a primary device, a secondary device, and a cloud device, and the method is performed by the primary device. The method includes: sending a first identity notification to the cloud device, where the first identity notification is a notification indicating that the primary device has a primary device identity; and uploading obtained first user information to the cloud device when determining that a communication status of the cloud device is normal, where the first user information is stored by the

cloud device and provided to the secondary device, and the first user information is to-be-backed-up information of user equipment that gets online from the primary device when the communication status of the cloud device is normal.

SUMMARY

[0004] In view of this, embodiments of this application provide a data backup method. In the method, during data backup, data is backed up at a granularity of a service, to avoid a business service error caused by data backup inconsistency, and ensure that service data backed up to a secondary cluster does not affect quality of a provided business service. This application further provides a corresponding data backup system, and a control device~~, a computing device, a computer-readable storage medium, and a computer program product~~.

[0005] According to a first aspect, embodiments of this application provide a data backup method. The method is applied to a data backup system including a primary cluster, a secondary cluster, and a control device, wherein the primary cluster includes a first component and a second component, the secondary cluster includes a third component and a fourth component, the third component is used as backup of the first component, the fourth component is used as backup of the second component, the first component and the second component store data in different formats, the third component and the fourth component store data in different formats. During specific implementation, the control device configures a first data backup policy for a first service based on: information about a plurality of data sets related to the first service and that is entered by a user; and a first moment, wherein the plurality of data sets related to the first service comprise a data set processed or stored by the first component in the primary cluster and a data set processed or stored by the second component in the primary cluster. The control device further controls, based on ~~a~~ the first data backup policy, the primary cluster or the secondary cluster to back up, to the secondary cluster, ~~a~~ the plurality of data sets related to ~~a~~ the first service that are in the primary cluster and that are at ~~a~~ the first moment, where the first data backup policy includes the information about the plurality of data sets related to the first service and the first moment.

[0006] In this way, data may be backed up between the primary cluster and the secondary cluster at a granularity of a service, so that the plurality of data sets related to the first service that are backed up to the secondary cluster may be consistent in a time dimension. In this way, when the primary cluster is faulty, the secondary cluster may restore service running based on service data in a same time period, to avoid a problem that an error occurs when the data backup system provides a business service because the service data backed up to the secondary cluster is inconsistent in a time dimension.

This may improve reliability of storing the service data for a user by the data backup system, and improve business service quality.

[0007] In a possible implementation, when the control device controls, based on a first data backup policy, the primary cluster or the secondary cluster to back up, to the secondary cluster, a plurality of data sets related to a first service that are in the primary cluster and that are at a first moment. Specifically, the control device may send a first instruction to the primary cluster, to instruct the primary cluster to send, to the secondary cluster, data corresponding to snapshots of the plurality of data sets related to the first service that are at the first moment. Alternatively, the control device sends a second instruction to the secondary cluster, to instruct the secondary cluster to replicate, from the primary cluster, data corresponding to snapshots of the plurality of data sets related to the first service that are at the first moment and that are in the primary cluster. In this way, the control device may control, by sending an instruction to the primary cluster or the secondary cluster, the primary cluster or the secondary cluster to implement a data backup process based on the snapshots.

[0008] In a possible implementation, before sending a first instruction to the primary cluster or sending a second instruction to the secondary cluster, the control device may first send a third instruction to the primary cluster, where the third instruction includes the information about the plurality of data sets related to the first service and the first moment, and the third instruction instructs the primary cluster to obtain the snapshots of the plurality of data sets related to the first service that are at the first moment. In this way, the primary cluster or the secondary cluster may subsequently back up, to the secondary cluster based on the snapshot corresponding to the first moment, the plurality of data sets related to the first service that are in the primary cluster, and the plurality of data sets that are backed up to the secondary cluster are all the plurality of data sets related to the first service that are in the primary cluster and that are at the first moment. In a manner of taking snapshots of the plurality of data sets related to the first service that are at the first moment, the data set at the first moment may be obtained and backed up more accurately, to avoid a problem of inconsistent data backup time caused by a communication delay.

[0009] In a possible implementation, the control device may further send a fourth instruction to the primary cluster, where the fourth instruction instructs the primary cluster to synchronize user data to the secondary cluster; or the control device may obtain user data stored in the primary cluster and the secondary cluster, and adjust, based on the user data stored in the primary cluster, the user data stored in the secondary cluster, so that the user data stored in the primary cluster is consistent with the user data stored in the secondary cluster. In this way, when the secondary cluster takes over a service on the primary cluster, the secondary cluster may provide a corresponding business service for a user based on the user data backed up to the secondary cluster, so that operation and maintenance

personnel do not need to manually configure the user data on the secondary cluster. In this way, not only operation and maintenance costs of the operation and maintenance personnel may be reduced, but also a recovery time objective of the data backup system may be effectively reduced.

[0010] For example, the user data may be, for example, at least one of a user identifier, user permission, or a tenant identifier, or may be other user-related data.

[0011] In a possible implementation, the control device may not only configure a first data backup policy for a first service, but also configure a second data backup policy for a second service, where the second data backup policy includes information about a plurality of data sets related to the second service and a second moment; and then, the control device controls, based on the second data backup policy, the primary cluster or the secondary cluster to back up, to the secondary cluster, the plurality of data sets related to the second service that are in the primary cluster and that are at the second moment. The second service and the first service belong to different services, and may be specifically different services belonging to a same user, or may be different services belonging to different users, or the like. In this way, the data backup system may implement data backup based on a service granularity for a plurality of different services, to support high-quality services of the plurality of services.

[0012] In a possible implementation, the plurality of data sets related to the first service include a data set processed or stored by a first component in the primary cluster and a data set processed or stored by a second component in the primary cluster. For example, the first component and the second component may be configured to be encapsulated into different formats, or the first component and the second component have different data processing performance. In this way, different data sets that belong to a same service and that are processed or stored by different components in the primary cluster at the first moment may be backed up to the secondary cluster.

[0013] In addition, the plurality of data sets related to the second service that are backed up to the secondary cluster in the data backup system may include a data set processed or stored by a first component in the primary cluster, a data set processed or stored by a second component in the primary cluster, a data set stored or processed by a third component in the primary cluster, and the like. Components that process or store data sets of different services may be different.

[0014] In a possible implementation, the control device includes a primary client and a secondary client. The primary client is configured to detect first status information of the primary cluster, and the secondary client is configured to detect second status information of the secondary cluster. In this case, the control device may further obtain the first status information obtained through detection of the primary client and the second status information obtained through detection of the secondary client. In addition, when the first status information indicates that the primary cluster is a secondary

identity or the cluster fails (for example, the primary cluster fails due to a fault), and when the second status information indicates that the secondary cluster is a primary identity, the control device determines that the secondary client is a client accessed by an application. In this way, when the primary/secondary identity of the primary cluster and the primary/secondary identity of the secondary cluster are reversed, the control device may automatically switch the client accessing the cluster, so that the operation and maintenance personnel do not need to perform manual switching.

[0015] For example, before the primary cluster is faulty, the first status information obtained by the control device may indicate that the primary cluster is the primary identity, and the second status information obtained by the control device may indicate that the secondary cluster is the secondary identity.

[0016] In a possible implementation, the control device may further prompt the user with information indicating that the primary cluster is faulty, so that the user determines, based on the prompt, that the primary cluster is faulty, and the control device may adjust the identity of the secondary cluster from the secondary identity to the primary identity in response to an identity adjustment operation performed by the user for the secondary cluster. In this way, identity reversal is performed by adjusting the primary cluster and the secondary cluster through a manual operation, so that abnormal switching of the primary/secondary identity of the primary cluster and the primary/secondary identity of the secondary cluster that is caused by a program running error in the data backup system may be avoided as much as possible.

[0017] In a possible implementation, the control device and the primary cluster are deployed in an isolated manner. For example, the control device and the secondary cluster may be jointly deployed at a secondary site, and the primary cluster is deployed at a primary site. Because the control device and the primary cluster are deployed in an isolated manner, when the primary cluster is faulty, the control device is not faulty, and the primary cluster may be switched to the control device when a fault occurs.

[0018] In a possible implementation, a same clock source is set in the control device, the primary cluster, and the secondary cluster. In this way, a moment at which the control device controls the primary cluster or the secondary cluster to perform data backup is consistent with a moment at which the primary cluster or the secondary cluster actually performs backup, to avoid a data backup error caused by inconsistent clock sources. Time consistency of data backup is improved.

[0019] In a possible implementation, the primary cluster and/or the secondary cluster include/includes a cluster constructed based on a hadoop architecture.

[0020] According to a second aspect, this application provides a data backup method. The method is applied to a data backup system, and the data backup system includes a primary cluster, a secondary

cluster, and a control device, wherein the primary cluster includes a first component and a second component, the secondary cluster includes a third component and a fourth component, the third component is used as backup of the first component, the fourth component is used as backup of the second component, the first component and the second component store data in different formats, the third component and the fourth component store data in different formats. The control device configures, a first data backup policy for a first service based on: information about a plurality of data sets related to the first service and that is entered by a user; and a first moment, wherein the plurality of data sets related to the first service comprise a data set processed or stored by the first component in the primary cluster and a data set processed or stored by the second component in the primary cluster. During specific implementation, the primary cluster obtains an instruction delivered by the control device, where the instruction summary includes information about a plurality of data sets related to a first service and a first moment, so that the primary cluster backs up, to the secondary cluster based on the instruction, the plurality of data sets related to the first service that are in the primary cluster and that are at the first moment, and synchronizing, by the primary cluster, user data to the secondary cluster. In this way, data may be backed up between the primary cluster and the secondary cluster at a granularity of a service, so that the plurality of data sets related to the first service that are backed up to the secondary cluster may be consistent in a time dimension.

[0021] In a possible implementation, when the primary cluster backs up, to the secondary cluster based on the instruction, the plurality of data sets related to the first service that are in the primary cluster and that are at the first moment, and specifically, may obtain, based on the information about the plurality of data sets related to the first service and the first moment, snapshots of the plurality of data sets related to the first service that are at the first moment and that are in the primary cluster. The primary cluster sends, data (namely, the plurality of data sets related to the first service) corresponding to the snapshots to the secondary cluster based on the snapshots, and backs up the plurality of data sets related to the first service that are in the primary cluster and that are at the first moment to the secondary cluster.

[0022] In a possible implementation, the primary cluster may further back up the user data to the secondary cluster. In this way, when the secondary cluster takes over a service on the primary cluster, the secondary cluster may provide a corresponding business service for a user based on the user data backed up to the secondary cluster, so that operation and maintenance personnel do not need to manually configure the user data on the secondary cluster. In this way, not only operation and maintenance costs of the operation and maintenance personnel may be reduced, but also a recovery time objective of the data backup system may be effectively reduced.

[0023] For example, the user data may be, for example, at least one of a user identifier, user

permission, or a tenant identifier, or may be other user-related data.

[0024] In a possible implementation, the primary cluster and/or the secondary cluster include/includes a cluster constructed based on a hadoop architecture.

[0025] ~~According to a third aspect, this application provides a control device, where the control device is located in a data backup system, the data backup system further includes a primary cluster and a secondary cluster, and the control device includes: a control module, configured to control, based on a first data backup policy, the primary cluster or the secondary cluster to back up, to the secondary cluster, a plurality of data sets related to a first service that are in the primary cluster and that are at a first moment, where the first data backup policy includes information about the plurality of data sets related to the first service and the first moment.~~

[0026] ~~In a possible implementation, the control module is specifically configured to: send a first instruction to the primary cluster, to instruct the primary cluster to send, to the secondary cluster, data corresponding to snapshots of the plurality of data sets related to the first service that are at the first moment, or send a second instruction to the secondary cluster, to instruct the secondary cluster to replicate, from the primary cluster, data corresponding to snapshots of the plurality of data sets related to the first service that are at the first moment and that are in the primary cluster.~~

[0027] ~~In a possible implementation, the control device further includes: a communication module, configured to: before the control device sends a first instruction to the primary cluster or the control device sends a second instruction to the secondary cluster, send a third instruction to the primary cluster, where the third instruction includes the information about the plurality of data sets related to the first service and the first moment, and the third instruction instructs the primary cluster to obtain the snapshots of the plurality of data sets related to the first service that are at the first moment.~~

[0028] ~~In a possible implementation, the control device further includes: a communication module, configured to send a fourth instruction to the primary cluster, where the fourth instruction instructs the primary cluster to synchronize user data to the secondary cluster; or the control module, further configured to: obtain, user data stored in the primary cluster and the secondary cluster, and adjust, based on the user data stored in the primary cluster, the user data stored in the secondary cluster.~~

[0029] ~~In a possible implementation, the control device further includes a configuration module, configured to configure the first data backup policy for the first service based on the information that is about the plurality of data sets related to the first service and that is entered by a user and the first moment.~~

[0030] ~~In a possible implementation, the control device further includes a configuration module, configured to configure a second data backup policy for a second service, where the second data~~

backup policy includes information about a plurality of data sets related to the second service and a second moment; and the control module is further configured to control, based on the second data backup policy, the primary cluster or the secondary cluster to back up, to the secondary cluster, the plurality of data sets related to the second service that are in the primary cluster and that are at the second moment.

[0031] In a possible implementation, the plurality of data sets related to the first service include a data set processed or stored by a first component in the primary cluster and a data set processed or stored by a second component in the primary cluster.

[0032] In a possible implementation, the control device includes a primary client and a secondary client, the primary client is configured to detect first status information of the primary cluster, the secondary client is configured to detect second status information of the secondary cluster, and the control device further includes: a communication module, configured to obtain the first status information obtained through detection of the primary client and the second status information obtained through detection of the secondary client; and a determining module, configured to: when the first status information indicates that the primary cluster is a secondary identity or the cluster fails, and the second status information indicates that the secondary cluster is a primary identity, determine that the secondary client is a client accessed by an application.

[0033] In a possible implementation, the control device further includes a prompting module and an adjustment module, where the prompting module is configured to prompt the user with information indicating that the primary cluster is faulty; and the adjustment module is configured to adjust an identity of the secondary cluster from the secondary identity to the primary identity in response to an identity adjustment operation of the user for the secondary cluster.

[0034] In a possible implementation, the control device is deployed in an isolated manner from the primary cluster.

[0035] In a possible implementation, a same clock source is set in the control device, the primary cluster, and the secondary cluster.

[0036] In a possible implementation, the primary cluster and/or the secondary cluster include/includes a cluster constructed based on a hadoop architecture.

[0037] According to a fourth aspect, this application provides a primary cluster, where the primary cluster is located in a data backup system, the data backup system further includes a secondary cluster and a control device, and the primary cluster includes: a communication module, configured to obtain an instruction delivered by the control device, where the instruction includes information about a plurality of data sets related to a first service and a first moment; and a backup module, configured to back up to the secondary cluster based on the instruction, the plurality of data

sets related to the first service that are in the primary cluster and that are at the first moment.

[0038] In a possible implementation, the backup module is specifically configured to: obtain, based on the information about the plurality of data sets related to the first service and the first moment, snapshots of the plurality of data sets related to the first service that are in the primary cluster and that are at the first moment; and send data corresponding to the snapshots to the secondary cluster based on the snapshots.

[0039] In a possible implementation, the backup module is further configured to synchronize the user data to the secondary cluster.

[0040] In a possible implementation, the primary cluster and/or the secondary cluster include/includes a cluster constructed based on a hadoop architecture.

[0041] According to a fifth third aspect, this application provides a data backup system. The data backup system includes a control device, a primary cluster, and a secondary cluster. The control device is configured to perform the data backup method according to the first aspect or any implementation of the first aspect, the primary cluster is configured to perform the data backup method according to the second aspect or any implementation of the second aspect, and the secondary cluster is configured to obtain and store a data set backed up from the primary cluster.

[0042] According to a sixth fourth aspect, this application provides a control device, where the control device includes a processor and a memory. The processor is configured to execute instructions stored in the memory, and the control device is enabled to perform the data backup method according to the first aspect or any implementation of the first aspect.

[0043] According to a seventh aspect, this application provides a primary cluster, where the primary cluster includes at least one processor and at least one memory, and the at least one processor executes instructions stored in the at least one memory, to enable the primary cluster to perform the data backup method according to the second aspect or any implementation of the second aspect.

[0044] According to an eighth aspect, this application provides a computer-readable storage medium, where the computer-readable storage medium stores instructions; and when the instructions run on a computing device, the computing device is enabled to perform the data backup method according to the first aspect or any implementation of the first aspect.

[0045] According to a ninth aspect, this application provides a computer-readable storage medium, where the computer-readable storage medium stores instructions; and when the instructions run on at least one computing device, the at least one computing device is enabled to perform the data backup method according to the second aspect or any implementation of the second aspect.

[0046] According to a tenth aspect, this application provides a computer program product including instructions. When the computer program product runs on a computing device, the

~~computing device is enabled to perform the data backup method according to the first aspect or any implementation of the first aspect.~~

[0047] ~~According to an eleventh aspect, this application provides a computer program product including instructions. When the computer program product runs on at least one computing device, the at least one computing device is enabled to perform the data backup method according to the second aspect or any implementation of the second aspect.~~

[0048] Based on the implementations provided in the foregoing aspects, this application may be further combined to provide more implementations.

BRIEF DESCRIPTION OF DRAWINGS

[0049] To describe technical solutions in embodiments of this application more clearly, the following briefly introduces accompanying drawings used for describing the embodiments. It is clear that the accompanying drawings in the following description show merely some embodiments of this application, and a person of ordinary skill in the art may still derive other drawings from these accompanying drawings.

[0050] FIG. 1 is a schematic diagram of an architecture of a data backup system 100;

[0051] FIG. 2 is a schematic diagram of backing up service data in a data backup system 100;

[0052] FIG. 3 is a schematic diagram of an architecture of a data backup system 300 according to an embodiment of this application;

[0053] FIG. 4 is a schematic diagram of backing up service data in a data backup system 300 according to an embodiment of this application;

[0054] FIG. 5 is a schematic diagram of an architecture of a data backup system 300 according to an embodiment of this application;

[0055] FIG. 6 is a schematic diagram of a cluster pairing interface according to an embodiment of this application;

[0056] FIG. 7 is a schematic diagram of a policy configuration interface according to an embodiment of this application;