Notice

DESCRIPTION CN112380067A

A big data backup system and method based on metadata in Hadoop environment

[0001]

Technical Field

[0002]

The present invention relates to the technical field of big data data storage, and in particular to a big data backup system and method based on metadata in a Hadoop environment.

[0003]

Background Art

[0004]

The Hadoop architecture is currently the most widely used big data architecture in the world. As the application areas of big data become wider and wider, the security of big data has received more and more attention.
Although current big data technology can achieve redundancy and platform data backup of big data platforms through multi-layer redundancy of data blocks in distributed architecture, current software backup technology cannot support scenarios such as recovery from platform administrator errors, data recovery from software version changes or software

bugs, time point-based data recovery, rapid backup and recovery of selected key data, and early warning of business impacts of backup or recovery operations on existing big data clusters.

[0005]

Summary of the invention

[0006]

The present invention aims to provide a big data backup system and method based on metadata in a Hadoop environment, which fully utilizes the distributed and high I/O characteristics of big data, supports rapid primary and secondary backup of key data in a big data platform, and makes intelligent backup recommendations for backup strategies based on current and historical performance records while ensuring data information security.

[0007]

To achieve the above object, the present invention is implemented by adopting the following technical solutions:

[0008]

The present invention discloses a big data backup system and method based on metadata in a Hadoop environment, including a backup client, a backup server, a backup strategy intelligent management terminal, a big data cluster terminal, and a big data backup cluster terminal.

[0009]

Backup client: used to provide users with visual backup access and customized backup plans;

[0010]

Backup server: includes production metadata synchronizer, production metadata list, primary backup metadata list, and secondary backup metadata list;

[0011]

Intelligent backup strategy management: stores backup strategies and intelligently recommends time windows for data backup or recovery based on historical cluster performance data;

[0012]

Big data cluster: used for the collection, integration, storage and analysis of big data, and for storing and restoring the primary backup data specified by the backup client;

[0013]

Big data backup cluster: used to store and restore secondary backup data specified by the client.

[0014]

Preferably, the backup server encrypts and synchronizes the Editlog log to the production metadata list in real time through the log monitoring program of the backup namenode in the big data cluster.

[0015]

The present invention also discloses a big data backup method using the backup system, including primary data backup, secondary data backup, primary data recovery, and secondary data recovery;

[0016]

Level 1 data backup includes the following steps:

[0017]

S11, the backup client accesses the backup server through the decryptor to obtain the latest metadata list.

[0018]

S12: The user uses the backup client to select files that need to be backed up from the metadata list.

[0019]

S13: The backup server submits a data replication job application for the backup file to the big data cluster according to the file list of the first-level data backup.

[0020]

S14. The log monitoring program on the big data cluster side finds the Editlog log of the backup data, and uses the encryption algorithm to generate a temporary file of the first-level backup metadata list in the first-level backup element list on the backup server side.

[0021]

S15. When the log monitoring program on the big data cluster side finds that the big data cluster is backed up successfully, the temporary file of the first-level backup metadata list on the backup server side is merged with the first-level backup metadata file.

[0022]

If the backup fails, delete the temporary file of the first-level backup metadata list;

[0023]

Secondary data backup includes the following steps:

[0024]

S21, the backup client accesses the backup server to obtain the latest metadata list.

[0025]

S22: The user uses the backup client to select files that need to be backed up at the secondary level from the metadata list.

[0026]

S23, the big data backup cluster side reads the corresponding file from the big data cluster side and writes it into the big data backup cluster side according to the backup file requirements.

[0027]

S24. The log monitoring program on the big data backup cluster side finds the Editlog log of the backup data, and uses the encryption algorithm to generate a temporary file of the secondary backup metadata list in the secondary backup element list on the backup server side.

[0028]

S25. When the log monitoring program on the backup big data cluster side finds that the big data cluster backup is successful, the secondary backup metadata list temporary file on the backup server side is merged with the secondary backup metadata file.

[0029]

If the backup fails, delete the temporary file of the secondary backup metadata list;

[0030]

Level 1 data recovery includes the following steps:

[0031]

S31. The backup client obtains the "first-level backup metadata list" from the backup server through a decryption algorithm, and obtains metadata information of the file list to be restored.

[0032]

S32. Find the data files to be restored based on the metadata information in the big data cluster.

[0033]

S33. Copy the data files to be restored in the big data cluster.

[0034]

S34. Use the log monitoring program on the big data cluster to monitor the data recovery status and synchronize it to the backup server in real time;

[0035]

Secondary data recovery includes the following steps:

[0036]

S41, the backup client obtains the metadata location of the "secondary backup metadata list" and the file list to be restored from the backup server through a decryption algorithm.

[0037]

S42: extract relevant recovery data from the big data backup cluster end according to the metadata position of the file list, and send a data write request to the big data cluster end to write the data to be recovered into the big data cluster end.

[0038]

S43. Use the log monitoring program on the big data backup cluster to monitor the data recovery status and synchronize it to the backup server in real time.

[0039]

Preferably, it also includes intelligent data backup and recovery, the steps of which are:

[0040]

S51. When the user submits a backup strategy application on the backup strategy intelligent management terminal, the backup strategy intelligent management terminal retrieves the historical cluster performance data and estimates the resources (CPU, memory, disk I/O, etc.) that will be occupied by backing up or restoring data based on the backup file size and file quantity, and determines whether the backup or restore operation will affect the normal computing use of the existing cluster.

[0041]

S52. When the data backup time selected by the user is estimated to affect the normal use of the big data cluster, the intelligent backup strategy management terminal will extract the cluster performance data of the past month, filter out the time window with CPU or memory usage less than 80% and no disk I/O delay and the cluster resource usage status of the corresponding time window, and find a similar time window based on the resources and backup time requirements required for this backup, and recommend the backup window to the user.

[0042]

S53. When the user manually initiates a strategic backup or recovery process, the backup strategy intelligent management terminal can view the current big data cluster performance.

[0043]

When the current CPU or memory usage of the big data cluster is greater than 80% or there is a large I/O delay, the user is prompted whether to force data backup or recovery.

[0044]

Preferably, the encryption algorithms in step S14 and step S24 are both AES and RSA hybrid encryption.

[0045]

Beneficial effects of the present invention:

[0046]

1. The present invention fully utilizes the current architectural features of HDFS, and has little difficulty in transforming the existing production big data platform.

[0047]

2. The present invention utilizes the distributed architecture of HDFS and the strong I/O concurrency feature to achieve faster data backup and recovery speeds.

[0048]

3. The present invention utilizes the HDFS redundant backup mechanism, which has high reliability in backing up and restoring data.

[0049]

4. The present invention adopts a metadata index backup method, so the backup method is flexible and can support full backup, incremental backup, off-site backup and other methods.

[0050]

5. The present invention encrypts the backup metadata, thereby improving the security of the data.

[0051]

6. The present invention can provide intelligent early warning and recommendation for data backup time windows.

[0052]

BRIEF DESCRIPTION OF THE DRAWINGS

[0053]

FIG. 1 is a schematic diagram of the architecture of the present invention.

[0054]

DETAILED DESCRIPTION

[0055]

In order to make the purpose, technical solutions and advantages of the present invention more clearly understood, the present invention is further described in detail below in conjunction with the accompanying drawings.

[0056]

In the present invention:

[0057]

Level 1 data backup refers to data backup in the production big data cluster.

[0058]

Secondary data backup refers to data backup in the big data backup cluster.

[0059]

Disk I/O refers to disk input and/or output operations.

[0060]

HDFS refers to Distributed File System.

[0061]

As shown in FIG1 , the present invention includes a backup client, a backup server, a backup strategy intelligent management terminal, a big data cluster terminal, and a big data backup cluster terminal.

[0062]

Backup client: used to provide users with visual backup access and customized backup plans;

[0063]

Backup server: includes production metadata synchronizer, production metadata list, primary backup metadata list, and secondary backup metadata list;

[0064]

Intelligent backup strategy management: stores backup strategies and intelligently recommends time windows for data backup or recovery based on historical cluster performance data;

[0065]

Big data cluster: used for the collection, integration, storage and analysis of big data, and for storing and restoring the primary backup data specified by the backup client;

[0066]

Big data backup cluster: used to store and restore secondary backup data specified by the client.

[0067]

The backup server uses the log monitoring program of the backup namenode in the big data cluster to encrypt and synchronize the Editlog log to the production metadata list in real time.

[0068]

The big data backup method of the above backup system mainly includes primary data backup, secondary data backup, primary data recovery, and secondary data recovery;

[0069]

Level 1 data backup includes the following steps:

[0070]

S11, the backup client accesses the backup server through the decryptor to obtain the latest metadata list.

[0071]

S12: The user uses the backup client to select files that need to be backed up from the metadata list.

[0072]

S13: The backup server submits a data replication job application for the backup file to the big data cluster according to the file list of the first-level data backup.

[0073]

S14. The log monitoring program on the big data cluster side finds the Editlog log of the backup data, and uses the encryption algorithm to generate a temporary file of the first-level backup metadata list in the first-level backup element list on the backup server side.

[0074]

S15. When the log monitoring program on the big data cluster side finds that the big data cluster is backed up successfully, the temporary file of the first-level backup metadata list on the backup server side is merged with the first-level backup metadata file.

[0075]

If the backup fails, delete the temporary file of the first-level backup metadata list;

[0076]

Secondary data backup includes the following steps:

[0077]

S21, the backup client accesses the backup server to obtain the latest metadata list.

[0078]

S22: The user uses the backup client to select files that need to be backed up at the secondary level from the metadata list.

[0079]

S23, the big data backup cluster side reads the corresponding file from the big data cluster side and writes it into the big data backup cluster side according to the backup file requirements.

[0080]

S24. The log monitoring program on the big data backup cluster side finds the Editlog log of the backup data, and uses the encryption algorithm to generate a temporary file of the secondary backup metadata list in the secondary backup element list on the backup server side.

[0081]

S25. When the log monitoring program on the backup big data cluster side finds that the big data cluster backup is successful, the secondary backup metadata list temporary file on the backup server side is merged with the secondary backup metadata file.

[0082]

If the backup fails, delete the temporary file of the secondary backup metadata list;

[0083]

Level 1 data recovery includes the following steps:

[0084]

S31. The backup client obtains the "first-level backup metadata list" from the backup server through a decryption algorithm, and obtains metadata information of the file list to be restored.

[0085]

S32. Find the data files to be restored based on the metadata information in the big data cluster.

[0086]

S33. Copy the data files to be restored in the big data cluster.

[0087]

S34. Use the log monitoring program on the big data cluster to monitor the data recovery status and synchronize it to the backup server in real time;

[0088]

Secondary data recovery includes the following steps:

[0089]

S41, the backup client obtains the metadata location of the "secondary backup metadata list" and the file list to be restored from the backup server through a decryption algorithm.

[0090]

S42: extract relevant recovery data from the big data backup cluster end according to the metadata position of the file list, and send a data write request to the big data cluster end to write the data to be recovered into the big data cluster end.

[0091]

S43. Use the log monitoring program on the big data backup cluster to monitor the data recovery status and synchronize it to the backup server in real time.

[0092]

Smart data backup and recovery, the steps are:

[0093]

S51. When the user submits a backup strategy application on the backup strategy intelligent management terminal, the backup strategy intelligent management terminal retrieves the historical cluster performance data and estimates the resources (CPU, memory, disk I/O, etc.) that will be occupied by backing up or restoring data based on the backup file size and file quantity, and determines whether the backup or restore operation will affect the normal computing use of the existing cluster.

[0094]

S52. When the data backup time selected by the user is estimated to affect the normal use of the big data cluster, the intelligent backup strategy management terminal will extract the cluster performance data of the past month, filter out the time window with CPU or memory usage less than 80% and no disk I/O delay and the cluster resource usage status of the corresponding time window, and find a similar time window based on the resources and backup time requirements required for this backup, and recommend the backup window to the user.

[0095]

S53. When the user manually initiates a strategic backup or recovery process, the backup strategy intelligent management terminal can view the current big data cluster performance.

[0096]

When the current CPU or memory usage of the big data cluster is greater than 80% or there is a large I/O delay, the user is prompted whether to force data backup or recovery.

[0097]

Of course, the present invention may have many other embodiments. Without departing from the spirit and essence of the present invention, technicians familiar with the field may make various corresponding changes and deformations based on the present invention, but these corresponding changes and deformations should all fall within the scope of protection of the claims attached to the present invention.

Patent Translate

Powered by EPO and Google

Notice

This translation is machine-generated. It cannot be guaranteed that it is intelligible, accurate, complete, reliable or fit for specific purposes. Critical decisions, such as commercially relevant or financial decisions, should not be based on machine-translation output.

CLAIMS CN112380067A

1.

A big data backup system based on metadata in a Hadoop environment, characterized by comprising a backup client, a backup server, a backup strategy intelligent management terminal, a big data cluster terminal, and a big data backup cluster terminal.
Backup client: used to provide users with visual backup access and customized backup plans;
Backup server: includes production metadata synchronizer, production metadata list, primary backup metadata list, and secondary backup metadata list;
Intelligent backup strategy management: stores backup strategies and intelligently recommends time windows for data backup or recovery based on historical cluster performance data;
Big data cluster: used for the collection, integration, storage and analysis of big data, and for storing and restoring the primary backup data specified by the backup client;
Big data backup cluster: used to store and restore secondary backup data specified by the client.

2.

The backup system according to claim 1 is characterized in that the backup server encrypts and synchronizes the Editlog log to the production metadata list in real time through the log monitoring program of the backup namenode in the big data cluster.

3.

A big data backup method using the backup system of claim 2, characterized by: comprising primary data backup, secondary data backup, primary data recovery, and secondary data recovery;

Level 1 data backup includes the following steps:

S11, the backup client accesses the backup server through the decryptor to obtain the latest metadata list.

S12: The user uses the backup client to select files that need to be backed up from the metadata list.

S13: The backup server submits a data replication job application for the backup file to the big data cluster according to the file list of the first-level data backup.

S14. The log monitoring program on the big data cluster side finds the Editlog log of the backup data, and uses the encryption algorithm to generate a temporary file of the first-level backup metadata list in the first-level backup element list on the backup server side.

S15. When the log monitoring program on the big data cluster side finds that the big data cluster is backed up successfully, the temporary file of the first-level backup metadata list on the backup server side is merged with the first-level backup metadata file.

If the backup fails, delete the temporary file of the first-level backup metadata list;

Secondary data backup includes the following steps:

S21, the backup client accesses the backup server to obtain the latest metadata list.

S22: The user uses the backup client to select files that need to be backed up at the secondary level from the metadata list.

S23, the big data backup cluster side reads the corresponding file from the big data cluster side and writes it into the big data backup cluster side according to the backup file requirements.

S24. The log monitoring program on the big data backup cluster side finds the Editlog log of the backup data, and uses the encryption algorithm to generate a temporary file of the secondary backup metadata list in the secondary backup element list on the backup server side.

S25. When the log monitoring program on the backup big data cluster side finds that the big data cluster backup is successful, the secondary backup metadata list temporary file on the backup server side is merged with the secondary backup metadata file.

If the backup fails, delete the temporary file of the secondary backup metadata list;

Level 1 data recovery includes the following steps:

S31. The backup client obtains the "first-level backup metadata list" from the backup server through a decryption algorithm, and obtains metadata information of the file list to be restored.

S32. Find the data files to be restored based on the metadata information in the big data cluster.

S33. Copy the data files to be restored in the big data cluster.
S34. Use the log monitoring program on the big data cluster to monitor the data recovery status and synchronize it to the backup server in real time;
Secondary data recovery includes the following steps:
S41, the backup client obtains the metadata location of the "secondary backup metadata list" and the file list to be restored from the backup server through a decryption algorithm.
S42: extract relevant recovery data from the big data backup cluster end according to the metadata position of the file list, and send a data write request to the big data cluster end to write the data to be recovered into the big data cluster end.
S43. Use the log monitoring program on the big data backup cluster to monitor the data recovery status and synchronize it to the backup server in real time.

4.

The backup method according to claim 3 is characterized in that it also includes intelligent data backup and recovery, the steps of which are:
S51. When a user submits a backup strategy application on the backup strategy intelligent management terminal, the backup strategy intelligent management terminal retrieves historical cluster performance data and estimates the resources that will be occupied by backing up or restoring data based on the size and number of backup files, and determines whether the backup or restore operation will affect the normal computing use of the existing cluster.
S52. When the data backup time selected by the user is estimated to affect the normal use of the big data cluster, the intelligent backup strategy management terminal will extract the cluster performance data of the past month, filter out the time window with CPU or memory usage less than 80% and no disk I/O delay and the cluster resource usage status of the corresponding time window, and find a similar time window based on the resources and backup time requirements required for this backup, and recommend the backup window to the user.
S53. When the user manually initiates a strategic backup or recovery process, the backup strategy intelligent management terminal can view the current big data cluster performance. When the current CPU or memory usage of the big data cluster is greater than 80% or there is a large I/O delay, the user is prompted whether to force data backup or recovery.

5.

The backup method according to claim 3 is characterized in that the encryption algorithms in step S14 and step S24 are both AES and RSA hybrid encryption.