(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112380067 A
(43)申请公布日 2021.02.19

(21)申请号 202011375213.4

(22)申请日 2020.11.30

(71)申请人 四川大学华西医院
地址 610000 四川省成都市武侯区国学巷37号

(72)发明人 胡耀 李春漾 应志野 张超 殷晋

(74)专利代理机构 成都高远知识产权代理事务所(普通合伙) 51222
代理人 李安霞 谢一平

(51)Int.Cl.
G06F 11/14(2006.01)
G06F 21/60(2013.01)
G06F 21/64(2013.01)

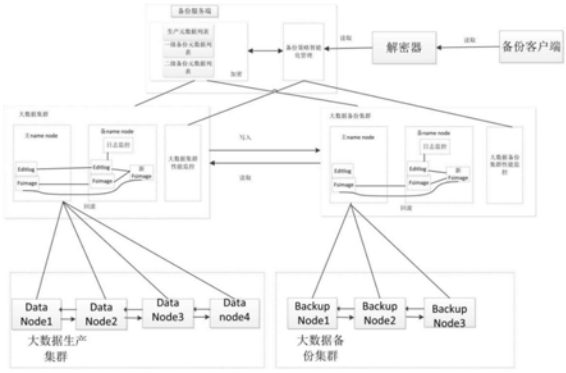
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种Hadoop环境下基于元数据的大数据备份系统及方法

(57)摘要

本发明公开一种Hadoop环境下基于元数据的大数据备份系统及方法，包括备份客户端、备份服务端、备份策略智能化管理端、大数据集群端、大数据备份集群端。本发明主要通过对Hadoop大数据环境的元数据架构的优化和调整，充利用大数据分布式，高I/O等特性，在保证数据信息安全的前提下，本发明可根据用户需求，支持大数据平台内关键数据快速进行集群内数据备份与恢复、不同集群间数据备份与恢复以及根据当前和历史记录对备份策略进行智能备份。



CN 112380067 A

1.一种Hadoop环境下基于元数据的大数据备份系统,其特征在于:包括备份客户端、备份服务端、备份策略智能化管理端、大数据集群端、大数据备份集群端,

备份客户端:用于为用户提供可视化备份访问、定制备份计划;

备份服务端:包括生产元数据同步器、生产元数据列表、一级备份元数据列表、二级备份元数据列表;

备份策略智能化管理端:对备份策略进行存储和根据集群历史性能数据智能推荐数据备份或恢复的时间窗口;

大数据集群端:用于大数据的采集、集成、存储与分析,存储和恢复由备份客户端指定的一级备份数据;

大数据备份集群端:用于存储和恢复客户端指定的二级备份数据。

2.根据权利要求1所述的备份系统,其特征在于:备份服务端通过大数据集群中备namenode的日志监控程序,实时加密同步Editlog日志至生产元数据列表中。

3.一种使用权利要求2所述备份系统的大数据备份方法,其特征在于:包括一级数据备份、二级数据备份、一级数据恢复、二级数据恢复;

一级数据备份包括以下步骤:

S11、备份客户端通过解密器访问备份服务端,获得最新的元数据清单列表,

S12、用户使用备份客户端从元数据清单列表中选择需要进行一级数据备份的文件,

S13、备份服务端根据一级数据备份的文件清单,向大数据集群端提交备份文件的数据复制作业申请,

S14、大数据集群端的日志监控程序发现备份数据的Editlog日志,并在备份服务端的一级备份元素列表中使用加密运算法生成一级备份元数据列表临时文件,

S15、当大数据集群端的日志监控程序发现大数据集群备份成功后,备份服务端的一级备份元数据列表临时文件与一级备份元数据文件合并,

如备份失败,则删除一级备份元数据列表临时文件;

二级数据备份包括以下步骤:

S21、备份客户端访问备份服务端,获得最新的元数据清单列表,

S22、用户使用备份客户端从元数据清单列表中选择需要进行二级数据备份的文件,

S23、大数据备份集群端根据需备份文件需求,从大数据集群端读取相应的文件并写入大数据备份集群端中,

S24、大数据备份集群端的日志监控程序发现备份数据的Editlog日志,并在备份服务端的二级备份元素列表中使用加密运算法生成二级备份元数据列表临时文件,

S25、当备大数据备份集群端的日志监控程序发现大数据集群备份成功后,备份服务端的二级备份元数据列表临时文件与二级备份元数据文件合并,

如备份失败,则删除二级备份元数据列表临时文件;

一级数据恢复包括以下步骤:

S31、备份客户端通过解密算法,从备份服务端获取“一级备份元数据列表”清单,并获得需要恢复的文件列表的元数据信息,

S32、在大数据集群端中根据元数据信息,找到需恢复的数据文件,

S33、在大数据集群端中复制需恢复的数据文件,

S34、利用大数据集群端的日志监控程序监控数据恢复状态，并实时同步至备份服务端上；

二级数据恢复包括以下步骤：

S41、备份客户端通过解密算法，从备份服务端获取“二级备份元数据列表”清单和需要恢复的文件列表的元数据位置，

S42、根据文件列表的元数据位置，在大数据备份集群端中提取相关恢复数据，并向大数据集群端发出写数据申请，将需恢复数据写入大数据集群端中，

S43、利用大数据备份集群端的日志监控程序监控数据恢复状态，并实时同步至备份服务端上。

4.根据权利要求3所述的备份方法，其特征在于：还包括智能数据备份与恢复，其步骤为：

S51、当用户在备份策略智能化管理端提交备份策略申请时，备份策略智能化管理端调取历史集群性能数据并根据备份文件大小、文件数量预估备份或恢复数据将会占用的资源，并且判断此次备份或恢复操作是否会影响现有集群正常的计算使用，

S52、当用户选择的数据备份时间预估会影响大数据集群端的正常使用时，备份策略智能化管理端会抽取近一月的集群性能数据，筛选出CPU或内存占用率小于80％且无磁盘I/O延迟的时间窗口和对应时间窗口的集群资源使用状态，并且根据此次备份需要占用资源和备份时间需求寻找相似的时间窗口，为用户推荐该备份窗口，

S53、当用户手动发起策略化备份或恢复进程时，备份策略智能化管理端可查看当前大数据集群性能情况，

当目前大数据集群端CPU或内存使用率大于80％或有较大I/O延迟时，则提示用户是否强制进行数据备份或恢复。

5.根据权利要求3所述的备份方法，其特征在于：步骤S14和步骤S24中的加密算法均为AES与RSA混合加密。

一种Hadoop环境下基于元数据的大数据备份系统及方法

技术领域

[0001] 本发明涉及大数据数据存储技术领域，尤其涉及一种Hadoop环境下基于元数据的大数据备份系统及方法。

背景技术

[0002] Hadoop架构目前是世界上应用最广泛的大数据架构，随着大数据应用领域越来越广，大数据的安全性越发的受到重视。虽然目前大数据技术通过分布式架构中数据块多层冗余的方式，已可以实现大数据平台的冗余与平台数据备份，但目前的软件备份技术无法支持平台管理员误操作恢复、软件版本变更或软件BUG的数据恢复，基于时间点的数据恢复和有选择的重点数据的快速备份与恢复，备份或恢复操作可能对现有大数据集群等业务影响预警等场景。

发明内容

[0003] 本发明旨在提供一种Hadoop环境下基于元数据的大数据备份系统及方法，充利用大数据分布式、高I/O等特性，在保证数据信息安全的前提下，支持大数据平台内关键数据进行快速一级和二级以及根据当前和历史性能记录对备份策略进行智能备份推荐。

[0004] 为达到上述目的，本发明是采用以下技术方案实现的：

[0005] 本发明公开一种Hadoop环境下基于元数据的大数据备份系统及方法，包括备份客户端、备份服务端、备份策略智能化管理端、大数据集群端、大数据备份集群端，

[0006] 备份客户端：用于为用户提供可视化备份访问、定制备份计划；

[0007] 备份服务端：包括生产元数据同步器、生产元数据列表、一级备份元数据列表、二级备份元数据列表；

[0008] 备份策略智能化管理端：对备份策略进行存储和根据集群历史性能数据智能推荐数据备份或恢复的时间窗口；

[0009] 大数据集群端：用于大数据的采集、集成、存储与分析，存储和恢复由备份客户端指定的一级备份数据；

[0010] 大数据备份集群端：用于存储和恢复客户端指定的二级备份数据。

[0011] 优选的，备份服务端通过大数据集群中备namenode的日志监控程序，实时加密同步Editlog日志至生产元数据列表中。

[0012] 本发明还公开使用上述备份系统的大数据备份方法，包括一级数据备份、二级数据备份、一级数据恢复、二级数据恢复；

[0013] 一级数据备份包括以下步骤：

[0014] S11、备份客户端通过解密器访问备份服务端，获得最新的元数据清单列表，

[0015] S12、用户使用备份客户端从元数据清单列表中选择需要进行一级数据备份的文件，

[0016] S13、备份服务端根据一级数据备份的文件清单，向大数据集群端提交备份文件的

数据复制作业申请，

[0017] S14、大数据集群端的日志监控程序发现备份数据的Editlog日志，并在备份服务端的一级备份元素列表中使用加密运算法生成一级备份元数据列表临时文件，

[0018] S15、当大数据集群端的日志监控程序发现大数据集群备份成功后，备份服务端的一级备份元数据列表临时文件与一级备份元数据文件合并，

[0019] 如备份失败，则删除一级备份元数据列表临时文件；

[0020] 二级数据备份包括以下步骤：

[0021] S21、备份客户端访问备份服务端，获得最新的元数据清单列表，

[0022] S22、用户使用备份客户端从元数据清单列表中选择需要进行二级数据备份的文件，

[0023] S23、大数据备份集群端根据需备份文件需求，从大数据集群端读取相应的文件并写入大数据备份集群端中，

[0024] S24、大数据备份集群端的日志监控程序发现备份数据的Editlog日志，并在备份服务端的二级备份元素列表中使用加密运算法生成二级备份元数据列表临时文件，

[0025] S25、当备大数据备份集群端的日志监控程序发现大数据集群备份成功后，备份服务端的二级备份元数据列表临时文件与二级备份元数据文件合并，

[0026] 如备份失败，则删除二级备份元数据列表临时文件；

[0027] 一级数据恢复包括以下步骤：

[0028] S31、备份客户端通过解密算法，从备份服务端获取“一级备份元数据列表”清单，并获得需要恢复的文件列表的元数据信息，

[0029] S32、在大数据集群端中根据元数据信息，找到需恢复的数据文件。

[0030] S33、在大数据集群端中复制需恢复的数据文件。

[0031] S34、利用大数据集群端的日志监控程序监控数据恢复状态，并实时同步至备份服务端上；

[0032] 二级数据恢复包括以下步骤：

[0033] S41、备份客户端通过解密算法，从备份服务端获取“二级备份元数据列表”清单和需要恢复的文件列表的元数据位置，

[0034] S42、根据文件列表的元数据位置，在大数据备份集群端中提取相关恢复数据，并向大数据集群端发出写数据申请，将需恢复数据写入大数据集群端中，

[0035] S43、利用大数据备份集群端的日志监控程序监控数据恢复状态，并实时同步至备份服务端上。

[0036] 优选的，还包括智能数据备份与恢复，其步骤为：

[0037] S51、当用户在备份策略智能化管理端提交备份策略申请时，备份策略智能化管理端调取历史集群性能数据并根据备份文件大小、文件数量预估备份或恢复数据将会占用的资源(CPU、内存、磁盘I/O等)，并且判断此次备份或恢复操作是否会影响现有集群正常的计算使用，

[0038] S52、当用户选择的数据备份时间预估会影响大数据集群端的正常使用时，备份策略智能化管理端会抽取近一月的集群性能数据，筛选出CPU或内存占用率小于80％且无磁盘I/O延迟的时间窗口和对应时间窗口的集群资源使用状态，并且根据此次备份需要占用

资源和备份时间需求寻找相似的时间窗口,为用户推荐该备份窗口,

[0039] S53、当用户手动发起策略化备份或恢复进程时,备份策略智能化管理端可查看当前大数据集群性能情况,

[0040] 当目前大数据集群端CPU或内存使用率大于80%或有较大I/O延迟时,则提示用户是否强制进行数据备份或恢复。

[0041] 优选的,步骤S14和步骤S24中的加密算法均为AES与RSA混合加密。

[0042] 本发明的有益效果:

[0043] 1、本发明充分的利用了HDFS现在架构特点,对现有生产大数所平台改造难度小。

[0044] 2、本发明利用HDFS的分布式架构,I/O并发力强的特点,数据备份和恢复速度较快。

[0045] 3、本发明利用HDFS冗余备份机制,此种方法备份和恢复数据可靠性强。

[0046] 4、本发明因采用元数据索引的备份的方式,所以备份方式灵活,可支持全备份,增量备份,异地备份等多种方式。

[0047] 5、本发明对备份元数据进行加密,从而提高了数据的安全性。

[0048] 6、本发明可对数据备份时间窗口进行智能预警和推荐。

附图说明

[0049] 图1为本发明的架构示意图。

具体实施方式

[0050] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图,对本发明进行进一步详细说明。

[0051] 本发明中:

[0052] 一级数据备份是指数据在生产大数据集群端中备份,

[0053] 二级数据备份是指数据在大数据备份集群端中备份,

[0054] 磁盘I/O是指磁盘的输入和/或输出操作,

[0055] HDFS是指分布式文件系统。

[0056] 如图1所示,本发明包括备份客户端、备份服务端、备份策略智能化管理端、大数据集群端、大数据备份集群端,

[0057] 备份客户端:用于为用户提供可视化备份访问、定制备份计划;

[0058] 备份服务端:包括生产元数据同步器、生产元数据列表、一级备份元数据列表、二级备份元数据列表;

[0059] 备份策略智能化管理端:对备份策略进行存储和根据集群历史性能数据智能推荐数据备份或恢复的时间窗口;

[0060] 大数据集群端:用于大数据的采集、集成、存储与分析,存储和恢复由备份客户端指定的一级备份数据;

[0061] 大数据备份集群端:用于存储和恢复客户端指定的二级备份数据。

[0062] 备份服务端通过大数据集群中备namenode的日志监控程序,实时加密同步Editlog日志至生产元数据列表中。

[0063] 上述备份系统的大数据备份方法主要包括一级数据备份、二级数据备份、一级数据恢复、二级数据恢复；

[0064] 一级数据备份包括以下步骤：

[0065] S11、备份客户端通过解密器访问备份服务端，获得最新的元数据清单列表，

[0066] S12、用户使用备份客户端从元数据清单列表中选择需要进行一级数据备份的文件，

[0067] S13、备份服务端根据一级数据备份的文件清单，向大数据集群端提交备份文件的数据复制作业申请，

[0068] S14、大数据集群端的日志监控程序发现备份数据的Editlog日志，并在备份服务端的一级备份元素列表中使用加密运算法生成一级备份元数据列表临时文件，

[0069] S15、当大数据集群端的日志监控程序发现大数据集群备份成功后，备份服务端的一级备份元数据列表临时文件与一级备份元数据文件合并，

[0070] 如备份失败，则删除一级备份元数据列表临时文件；

[0071] 二级数据备份包括以下步骤：

[0072] S21、备份客户端访问备份服务端，获得最新的元数据清单列表，

[0073] S22、用户使用备份客户端从元数据清单列表中选择需要进行二级数据备份的文件，

[0074] S23、大数据备份集群端根据需备份文件需求，从大数据集群端读取相应的文件并写入大数据备份集群端中，

[0075] S24、大数据备份集群端的日志监控程序发现备份数据的Editlog日志，并在备份服务端的二级备份元素列表中使用加密运算法生成二级备份元数据列表临时文件，

[0076] S25、当备大数据备份集群端的日志监控程序发现大数据集群备份成功后，备份服务端的二级备份元数据列表临时文件与二级备份元数据文件合并，

[0077] 如备份失败，则删除二级备份元数据列表临时文件；

[0078] 一级数据恢复包括以下步骤：

[0079] S31、备份客户端通过解密算法，从备份服务端获取“一级备份元数据列表”清单，并获得需要恢复的文件列表的元数据信息，

[0080] S32、在大数据集群端中根据元数据信息，找到需恢复的数据文件。

[0081] S33、在大数据集群端中复制需恢复的数据文件。

[0082] S34、利用大数据集群端的日志监控程序监控数据恢复状态，并实时同步至备份服务端上；

[0083] 二级数据恢复包括以下步骤：

[0084] S41、备份客户端通过解密算法，从备份服务端获取“二级备份元数据列表”清单和需要恢复的文件列表的元数据位置，

[0085] S42、根据文件列表的元数据位置，在大数据备份集群端中提取相关恢复数据，并向大数据集群端发出写数据申请，将需恢复数据写入大数据集群端中，

[0086] S43、利用大数据备份集群端的日志监控程序监控数据恢复状态，并实时同步至备份服务端上。

[0087] 智能数据备份与恢复，其步骤为：

[0088] S51、当用户在备份策略智能化管理端提交备份策略申请时，备份策略智能化管理端调取历史集群性能数据并根据备份文件大小、文件数量预估备份或恢复数据将会占用的资源(CPU、内存、磁盘I/O等)，并且判断此次备份或恢复操作是否会影响现有集群正常的计算使用，

[0089] S52、当用户选择的数据备份时间预估会影响大数据集群端的正常使用时，备份策略智能化管理端会抽取近一月的集群性能数据，筛选出CPU或内存占用率小于80％且无磁盘I/O延迟的时间窗口和对应时间窗口的集群资源使用状态，并且根据此次备份需要占用资源和备份时间需求寻找相似的时间窗口，为用户推荐该备份窗口，

[0090] S53、当用户手动发起策略化备份或恢复进程时，备份策略智能化管理端可查看当前大数据集群性能情况，

[0091] 当目前大数据集群端CPU或内存使用率大于80％或有较大I/O延迟时，则提示用户是否强制进行数据备份或恢复。

[0092] 当然，本发明还可有其它多种实施例，在不背离本发明精神及其实质的情况下，熟悉本领域的技术人员可根据本发明作出各种相应的改变和变形，但这些相应的改变和变形都应属于本发明所附的权利要求的保护范围。

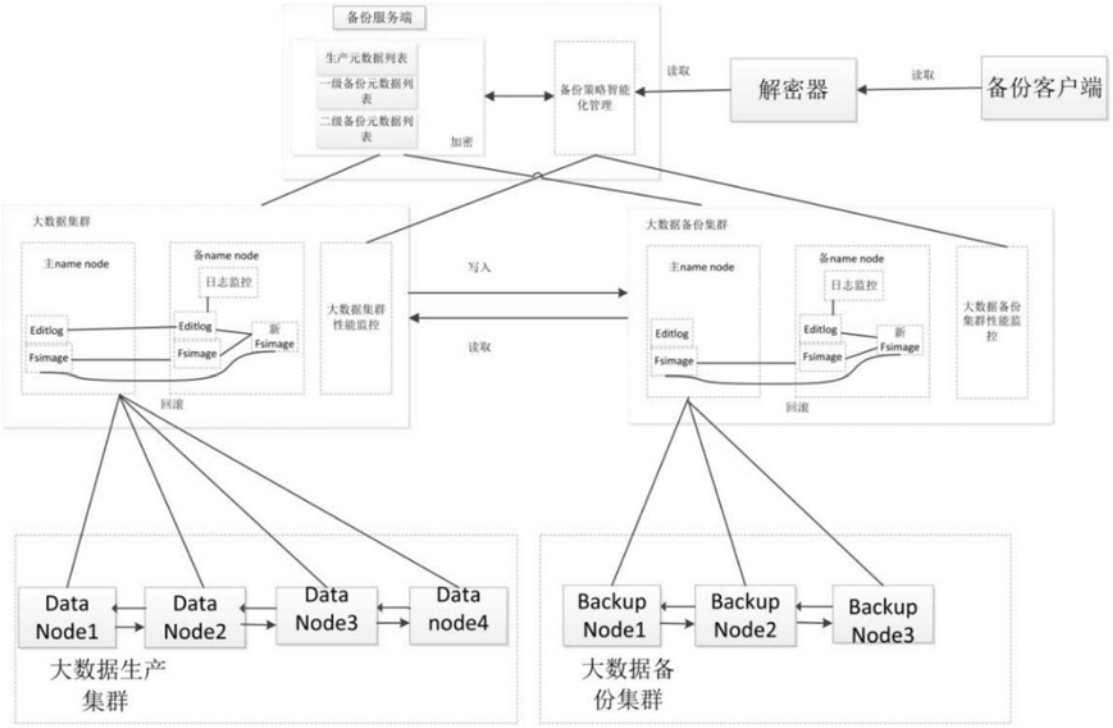

备份服务端
生产元数据列表
一级备份元数据列表
二级备份元数据列表
加密
备份策略智能化管理
读取
解密器
读取
备份客户端
大数据集群
主name node
备name node
日志监控
Editlog
Fsimage
Editlog
Fsimage
新Fsimage
大数据集群性能监控
写入
读取
回滚
大数据备份集群
主name node
备name node
日志监控
Editlog
Fsimage
Editlog
Fsimage
新Fsimage
大数据备份集群性能监控
回滚
Data Node1
Data Node2
Data Node3
Data node4
大数据生产集群
Backup Node1
Backup Node2
Backup Node3
大数据备份集群

图1