(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108268614 B
(45)授权公告日 2020.08.18

(21)申请号 201711486203.6

(22)申请日 2017.12.29

(65)同一申请的已公布的文献号
申请公布号 CN 108268614 A

(43)申请公布日 2018.07.10

(73)专利权人 郑州轻工业学院
地址 450000 河南省郑州市金水区东风路5号

(72)发明人 殷君茹 王华 孟颖辉 黄伟 朱付保

(74)专利代理机构 深圳市威世博知识产权代理事务所(普通合伙) 44280
代理人 李庆波

(51)Int.Cl.
G06F 16/182(2019.01)
G06F 16/13(2019.01)
G06F 16/11(2019.01)

(56)对比文件
CN 103678691 A,2014.03.26
CN 105677826 A,2016.06.15
US 9460147 B1,2016.10.04
邢乐乐.面向海量森林资源信息的云计算作业调度算法.《中国优秀硕士学位论文数据库信息科技辑》.2014,(第03期),I138-24.

审查员 李萌

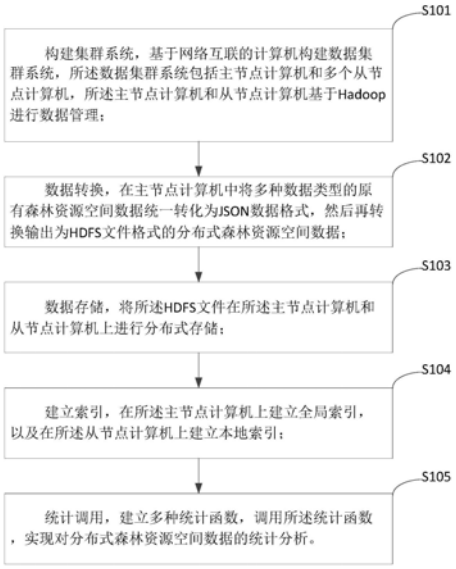
权利要求书1页 说明书9页 附图4页

(54)发明名称

一种森林资源空间数据的分布式管理方法

(57)摘要

本发明公开了一种森林资源空间数据的分布式管理方法,属于大数据计算领域,该方法包括步骤构建集群系统、数据转换、数据存储、建立索引和统计调用。通过该方法能够把多种类型的森林资源空间数据统一转化为适用于Hadoop集群系统的HDFS文件进行存储,并通过hive、spatialHadoop等软件工具实现对森林资源空间数据的高效统计,具有兼容性强、效率高以及技术开发成本低等优势。



CN 108268614 B

1.一种森林资源空间数据的分布式管理方法,其特征在于,

构建集群系统,基于网络互联的计算机构建数据集群系统,所述数据集群系统包括主节点计算机和多个从节点计算机,所述主节点计算机和从节点计算机基于Hadoop进行数据管理;

数据转换,将多种数据类型的原有森林资源空间数据统一转化为JSON数据格式,然后再转换输出为HDFS文件格式的分布式森林资源空间数据;所述原有森林资源空间数据的类型包括森林资源调查数据、概要统计数据和林地落界数据;所述林地落界数据包含县、省、国家级各级尺度或不同比例尺数据,又通过对不同种类、内容、来源和用户的业务数据的转化、提取和挖掘,准确反映林地资源业务数据之间的关联关系;数据类型包括面状矢量数据,既包括反映森林资源现状和变化的属性信息,也包括反映空间特征的信息;森林资源空间数据的数据类型包括结构化数据和非结构化数据;

基于Hive软件使用SerDes将结构化和非结构化的森林资源空间数据转换为JSON数据格式,进一步建立数据表,该数据表把JSON数据格式的森林资源空间数据的名称和地理信息映射为该数据表中的列名和边界形状,得到HDFS文件格式的数据表;

数据存储,将所述HDFS文件在所述主节点计算机和从节点计算机上进行分布式存储;

建立索引,在所述主节点计算机上建立全局索引,以及在所述从节点计算机上建立本地索引;

统计调用,建立多种统计函数,调用所述统计函数,实现对分布式森林资源空间数据的统计分析。

2.根据权利要求1所述的森林资源空间数据的分布式管理方法,其特征在于,所述主节点计算机和所述从节点计算机之间,以及多个所述从节点计算机之间无密码验证登录。

3.根据权利要求2所述的森林资源空间数据的分布式管理方法,其特征在于,所述HDFS文件以大小相等的数据块进行存储。

4.根据权利要求3所述的森林资源空间数据的分布式管理方法,其特征在于,所述数据块的大小为128Mbit。

5.根据权利要求1所述的森林资源空间数据的分布式管理方法,其特征在于,在所述数据转换中,还包括对原有的森林资源空间数据库迁移到所述数据集群系统。

6.根据权利要求1所述的森林资源空间数据的分布式管理方法,其特征在于,在所述建立索引中,还包括在在建立本地索引和全局索引之前先建立分区。

7.根据权利要求6所述的森林资源空间数据的分布式管理方法,其特征在于,在所述建立分区中,包括计算分区数、确定分区边界和物理分区。

8.根据权利要求7所述的森林资源空间数据的分布式管理方法,其特征在于,在所述建立分区中,包括步骤:建立统计函数、根据业务需求编写HQL语句、把结果存入新建表格以及将统计结果可视化显示。

一种森林资源空间数据的分布式管理方法

技术领域

[0001] 本发明涉及大数据计算领域，尤其涉及一种森林资源空间数据的分布式管理方法。

背景技术

[0002] 森林资源空间数据是指森林资源信息与地理空间信息相结合的一种数据，具有数据量大、信息内容多等特点。特别是近些年，随着我国北斗定位系统在林业的广泛应用，森林资源空间数据的类型不断涌现，除了已有的结构化数据外，还有非结构化数据出现。

[0003] 当把这些数据量庞大的、类型差异化的森林资源空间数据进行统一存储管理时，就需要为这些数据提供有效的处理方法，以保证对不同结构类型的数据进行统一化的处理，同时也能够适应海量数据的网络化存储和调用需求，能够高效快捷的存储、调用、查询这些森林资源空间数据。

[0004] 为此，本发明提供用以解决上述问题的一种森林资源空间数据的分布式管理方法。

发明内容

[0005] 本发明主要解决的技术问题是提供一种森林资源空间数据的分布式管理方法，解决现有技术中因森林资源空间数据结构多样化而导致的难以统一进行存储和使用的问题。

[0006] 为解决上述技术问题，本发明采用的一种技术方案是提供一种森林资源空间数据的分布式管理方法，包括：构建集群系统，基于网络互联的计算机构建数据集群系统，所述数据集群系统包括主节点计算机和多个从节点计算机，所述主节点计算机和从节点计算机基于Hadoop进行数据管理；数据转换，在主节点计算机中将多种数据类型的原有森林资源空间数据统一转化为JSON数据格式，然后再转换输出为HDFS文件格式的分布式森林资源空间数据；数据存储，将所述HDFS文件在所述主节点计算机和从节点计算机上进行分布式存储；建立索引，在所述主节点计算机上建立全局索引，以及在所述从节点计算机上建立本地索引；统计调用，建立多种统计函数，调用所述统计函数，实现对分布式森林资源空间数据的统计分析。

[0007] 在本发明森林资源空间数据的分布式管理方法另一实施例中，所述主节点计算机和所述从节点计算机之间，以及多个所述从节点计算机之间无密码验证登录。

[0008] 在本发明森林资源空间数据的分布式管理方法另一实施例中，在所述数据转换中先把所述森林资源空间数据转换为JSON数据格式，再利用Hive软件建立数据表，然后把所述JSON数据格式的森林资源空间数据加载到所述数据表中，得到所述HDFS文件格式的数据表。

[0009] 在本发明森林资源空间数据的分布式管理方法另一实施例中，所述原有森林资源空间数据的类型包括森林资源调查数据、概要统计数据和林地落界数据。

[0010] 在本发明森林资源空间数据的分布式管理方法另一实施例中，所述HDFS文件以大

小相等的数据块进行存储。

[0011] 在本发明森林资源空间数据的分布式管理方法另一实施例中，所述数据块的大小为128Mbit。

[0012] 在本发明森林资源空间数据的分布式管理方法另一实施例中，在所述数据转换中，还包括对原有的森林资源空间数据库迁移到所述数据集群系统。

[0013] 在本发明森林资源空间数据的分布式管理方法另一实施例中，在所述建立索引中，还包括在在建立本地索引和全局索引之前先建立分区。

[0014] 在本发明森林资源空间数据的分布式管理方法另一实施例中，在所述建立分区中，包括计算分区数、确定分区边界和物理分区。

[0015] 在本发明森林资源空间数据的分布式管理方法另一实施例中，在所述建立分区中，包括步骤：建立统计函数、根据业务需求编写HQL语句、把结果存入新建表格以及将统计结果可视化显示。

[0016] 本发明的技术效果是：本发明实施例公开一种森林资源空间数据的分布式管理方法，该方法包括步骤构建集群系统、数据转换、数据存储、建立索引和统计调用。通过该方法能够把多种类型的森林资源空间数据统一转化为适用于Hadoop集群系统的HDFS文件进行存储，并通过hive、spatialHadoop等软件工具实现对森林资源空间数据的高效统计，具有兼容性强、效率高以及技术开发成本低等优势。

附图说明

[0017] 图1是根据本发明森林资源空间数据的分布式管理方法一实施例的流程图；

[0018] 图2是根据本发明森林资源空间数据的分布式管理方法另一实施例中的集群系统组成示意图；

[0019] 图3是根据本发明森林资源空间数据的分布式管理方法另一实施例中的无密码登录配置示意图；

[0020] 图4是根据本发明森林资源空间数据的分布式管理方法另一实施例中的无密码登录配置示意图；

[0021] 图5是根据本发明森林资源空间数据的分布式管理方法另一实施例中的无密码登录配置示意图；

[0022] 图6是根据本发明森林资源空间数据的分布式管理方法一实施例中的森林资源空间数据的类型示例图。

具体实施方式

[0023] 为了便于理解本发明，下面结合附图和具体实施例，对本发明进行更详细的说明。附图中给出了本发明的较佳的实施例。但是，本发明可以以许多不同的形式来实现，并不限于本说明书所描述的实施例。相反地，提供这些实施例的目的是使对本发明的公开内容的理解更加透彻全面。

[0024] 需要说明的是，除非另有定义，本说明书所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。

[0025] 在本发明的说明书中所使用的术语只是为了描述具体的实施例的目的，不是用于

限制本发明。本说明书所使用的术语“和/或”包括一个或多个相关的所列项目的任意的和所有的组合。

[0026] 图1显示了本发明基于模板的蒙版自动擦除方法一实施例的流程图。在图1中，包括：

[0027] 步骤S101：构建集群系统，基于网络互联的计算机构建数据集群系统，所述数据集群系统包括主节点计算机和多个从节点计算机，所述主节点计算机和从节点计算机基于Hadoop进行数据管理；

[0028] 步骤S102：数据转换，在主节点计算机中将多种类型的森林资源空间数据统一转化为JSON数据格式，然后再转换输出为HDFS文件格式的分布式森林资源空间数据；

[0029] 步骤S103：数据存储，将所述HDFS文件在所述主节点计算机和从节点计算机上进行分布式存储；

[0030] 步骤S104：建立索引，在所述主节点计算机上建立全局索引，以及在所述从节点计算机上建立本地索引；

[0031] 步骤S105：统计调用，建立多种统计函数，调用所述统计函数，实现对所述分布式森林资源空间数据的统计分析。

[0032] 以下对上述步骤进一步说明。

[0033] 首先，步骤S101是为了解决森林资源空间数据进行分布式管理的组织结构问题，就是需要建立一个基于网络互联的数据管理集群系统，而网络互联的主体设备就是计算机(包括用作服务器的计算机)。

[0034] 如图2所示，进一步给出了这种集群系统的示意组成图。图2中，包括主节点计算机11和多个从节点计算机12，这些计算机通过网络13互联，网络13既可以是计算机局域网，也可以是Internet网。

[0035] 优选的，图2中的主节点计算机11和多个从节点计算机12基于Hadoop进行数据管理。这里，Hadoop是一个能够对大量数据进行分布式处理的软件框架，因此在主节点计算机11和多个从节点计算机12上安装有相应的Hadoop工具包软件，如包括ZooKeeper软件、Hbase软件、Hive软件、spatialHadoop软件，以及对这些软件进行必要的配置操作。其中，ZooKeeper软件是一个分布式的应用程序协调服务软件，是用于提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务等；Hive软件是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的SQL(Structured Query Language)查询功能，可以将SQL语句转换为MapReduce任务进行运行。HBase(Hadoop Database)软件是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统。

[0036] 优选的，在主节点计算机11和所有从节点计算机12之间实现无密码验证登录。

[0037] 图3反映了主节点计算机无密码登陆所有从节点计算机的示意图，图3中的MainCP代表主节点计算机，对应的IP地址是25.21.38.2，WorkerCP01、WorkerCP02、WorkerCP02分别代表从节点计算机，分别对应的IP地址是25.21.38.7、25.21.38.5、25.21.38.9。具体实现过程可以参考以下实施例：

[0038] 1.主节点计算机上生成密码对。

[0039] 以Hadoop用户身份登陆，在MainCP节点上执行以下命令：

[0040] SSH-Keygen-t rsa-P” #生成无密码密钥对

[0041] 2.查看“/home/Hadoop”下是否有“.ssh”文件夹,且“.ssh”文件夹是否有两个刚生成的无密码密钥对。

[0042] [Hadoop@MainCP～]$ll-a|grep.ssh

[0043] 3.追加id_rsa.pub到授权的密钥Key里面。

[0044] cat～/.ssh/id_rsa.pub>>～/.ssh/authorized_keys

[0045] 4.修改文件“authorized_keys”权限

[0046] chmod 600～/.ssh/authorized_keys

[0047] 5.设置无密码验证配置

[0048] 用root用户登陆,修改SSH配置文件,该文件位于“/etc/ssh/sshd_config”。修改内容如下:

[0049]

RSAAuthentication yes	#启用RSA认证
PubkeyAuthentication yes	#启用公钥私钥配对认证方式
AuthorizedKeysFile.ssh/authorized	#公钥文件路径(和上面生成的文件同)

[0050] 6.重启无密码验证服务,使设置生效。

[0051] service sshd restart

[0052] 7.退出root登陆,使用Hadoop普通用户验证是否成功。

[0053] ssh localhost

[0054] 8.以IP为“25.21.38.7”的从节点计算机为例,进行从主节点计算机到从节点计算机的配置。

[0055] a.把公钥复制到所有的从节点计算机上,以25.21.38.7为例,使用如下命令。

[0056] scp～/.ssh/id_rsa.pub Hadoop@25.21.38.7:～/

[0057] b.以Hadoop用户身份登陆,查看“/home/Hadoop”下是否存在这个文件。

[0058] c.在“/home/Hadoop”下创建“.ssh”文件夹。(备注:如以存在,则无需创建。)

[0059] 命令如下:

[0060] mkdir～/.ssh

[0061] 修改文件夹“.ssh”的权限:

[0062] chmod 700～/.ssh

[0063] d.追加到授权文件“authorized_keys”。

[0064] cat～/id_rsa.pub>>～/.ssh/authorized_keys

[0065] chmod 600～/.ssh/authorized_keys

[0066] e.用root用户修改“/etc/ssh/sshd_config”

[0067] 具体步骤参看前面MainCP的“设置无密码验证配置”。分两步:一修改配置文件;二重启服务。

[0068] f.用MainCP使用SSH无密码登陆25.21.38.7

[0069] SSH 25.21.38.7

[0070] g.删除“/home/Hadoop/”目录下的“id_rsa.pub”文件。

[0071] rm-rf～/id_rsa.pub

[0072] 9.其他从节点计算机的配置,参考步骤8.

[0073] 通过该实施例可以看出，通过在主节点计算机和从节点计算机上设置密钥对的方式实现了无密码登录，因此在应用上表现为无密码登录，但实际上是通过构成集群系统的计算机之间设置互相认可的密钥实现了登录，因此，需要双方都要进行配置，如上述实施例中在主节点计算机上生成密码对，又把公钥复制到从节点计算机上。由此在保证安全性的同时，又提高了这些计算机互联互通互访问的效率。

[0074] 进一步的，图4反映了从节点计算机无密码登陆主节点计算机的示意图，图4中的MainCP代表主节点计算机，对应的IP地址是25.21.38.2，WorkerCP01、WorkerCP02、WorkerCP02分别代表从节点计算机，分别对应的IP地址是25.21.38.7、25.21.38.5、25.21.38.9。具体实现过程可以参考以下实施例说明(以IP地址是25.21.38.7的从节点计算机为例)：

[0075] 1.创建该从节点计算机的公钥和私钥，并把自己的公钥追加到“authorized_keys”中。以Hadoop用户身份登陆，使用如下命令：

[0076] ssh-keygen-t rsa-P”

[0077] cat～/.ssh/id_rsa.pub>>～/.ssh/authorized_keys

[0078] 2.复制“25.21.38.7”的公钥“id_rsa.pub”到“MainCP”的“/home/Hadoop”目录下，并追加到“MainCP”的“authorized_keys”中。

[0079] a.在“25.21.38.7”从节点计算机上操作

[0080] scp～/.ssh/id_rsa.pub Hadoop@25.21.38.7:～/

[0081] b.在“MainCP”主节点计算机操作

[0082] 以Hadoop用户身份登陆，用如下命令：

[0083] cat～/id_rsa.pub>>～/.ssh/authorized_keys

[0084] c.删除刚复制过来的“id_rsa.pub”文件。

[0085] rm～/id_rsa.pub

[0086] 3.其他从节点计算机操作步骤一样。

[0087] 进一步的，图5反映了从节点计算机之间无密码登陆的示意图，图5中的WorkerCP01、WorkerCP02、WorkerCP02分别代表从节点计算机，分别对应的IP地址是25.21.38.7、25.21.38.5、25.21.38.9。具体实现过程可以参考以下实施例说明(以IP地址是25.21.38.7和25.21.38.5的两个从节点计算机为例)：

[0088] 1.复制“25.21.38.7”的公钥“id_rsa.pub”到“25.21.38.5”的“/home/Hadoop”目录下，并追加到“25.21.38.5”的“authorized_keys”中。

[0089] a.在“25.21.38.7”从节点计算机上操作

[0090] scp～/.ssh/id_rsa.pub Hadoop@25.21.38.5:～/

[0091] b.在“25.21.38.5”从节点计算机上操作

[0092] 以Hadoop用户身份登陆，用如下命令：

[0093] cat～/id_rsa.pub>>～/.ssh/authorized_keys

[0094] c.删除刚复制过来的“id_rsa.pub”文件。

[0095] rm～/id_rsa.pub

[0096] 2.复制“25.21.38.5”的公钥“id_rsa.pub”到“25.21.38.7”的“/home/Hadoop”目录下，并追加到“25.21.38.7”的“authorized_keys”中。

[0097] a.在“25.21.38.5”从节点计算机上操作

[0098] scp～/.ssh/id_rsa.pub Hadoop@25.21.38.7:～/

[0099] b.在“25.21.38.7”从节点计算机上操作

[0100] 以Hadoop用户身份登陆,用如下命令:

[0101] cat～/id_rsa.pub>>～/.ssh/authorized_keys

[0102] c.删除刚复制过来的“id_rsa.pub”文件。

[0103] 3.以Hadoop身份加以验证。

[0104] 4.其他服务器之间的设置参看步骤1-3。

[0105] 由此可以实现在主节点计算机11和所有从节点计算机12之间进行无密码登录,在保证系统安全性的基础上有利于提高整个系统的运行效率。

[0106] 优选的,主节点计算机11在Hadoop所基于的HDFS(Hadoop Distributed File System)内部提供元数据服务,而从节点计算机12为HDFS提供存储块。优选的,这该集群系统中,HDFS文件以大小相等的数据块进行存储,例如,优选这种数据块的大小为128Mbit。

[0107] 进一步的,对于步骤S102,主要是解决对多种类型的森林资源空间数据进行统一转化的问题。

[0108] 对于森林资源空间数据而言,一方面是林业本身的业务数据,这些数据根据不同的尺度和业务应用,数据可大致分为两类:第一类数据是用于县、乡、国有林场这些部门采集使用的详细的、以二类小班数据为主的森林资源调查数据,如表1所示;第二类数据是满足省级和国家级部门拟定国家林业发展战略、中长期发展规划并组织实施的概要统计数据;第三类数据是要把林业数据与空间地理位置结合起来,即林地落界数据,如表2所示,这些林地落界数据所具有的数据尺度也是不同的,既包含县、省、国家级各级尺度或不同比例尺数据,同时,又通过对不同种类、内容、来源和用户的业务数据的转化、提取和挖掘,准确反映林地资源业务数据之间的关联关系,并为管理者对森林资源的全面掌握提供准确、全面的数据支撑。

[0109] 从数据类型来看包括面状矢量数据,既包括反映森林资源现状和变化的属性信息,如权属、地类、优势树种、面积等,也包括反映空间特征的信息,如空间数据类型、空间位置坐标等。

[0110] 表1 森林资源空间数据示例一

[0111]

字段	字段类型	字段大小	说明
SHENG	Text	2	省(区、市)
XIAN	Text	6	县(市、旗)
XIAO_BAN	Text	4	图斑(小班)
DI_MAO	Text	1	地貌
PO_XIANG	Text	1	坡向
PO_WEI	Text	1	坡位
PO_DU	Short Integer	5	坡度
KE_JI_DU	Text	1	交通区位
TU_RANG_LX	Text	20	土壤类型(名称)
TU_CENG_HD	Short Integer	5	土层厚度

LD_QS	Text	2	土地权属
LIN_ZHONG	Text	3	地类
LD_KD	Double	38	林带宽度
LD_CD	Double	38	林带长度

[0112] 优选的，图6进一步显示了我国的林业数据库的基本架构及数据类型。可以看出森林资源空间数据的数据类型内容很多，包括结构化数据和非结构化数据，仅从数据存储类型描述上，森林资源空间数据的数据类型就包括有字符串、整型、双精度等。

[0113] 为此，为了在图2所示的数据集群系统中存储和使用多种类型的森林资源空间数据，需要对这些数据进行转化处理。

[0114] 表2 森林资源空间数据示例二

[0115]

字段名	字段别名	字段序号	字段长度
LYRID	图层编号	1	10
LYRLABEL	图层名称	2	50
LYRTYPE	图层类型	3	5
VISIBLE	是否可视	4	4
ISOPTLYR	是否叠加图层	5	4
VISLYRS	VGST图层名称	6	50
LYRADD	所属地	7	20
LYRTHEM	图层专题名	8	10
LYRINDEX	图层序号	9	4
INFOID	图层配置信息编号	10	10
URL	图层URL路径	11	1073741822
ORIGINPOINT	起始点	8	100
EXTENT	图层初始范围	9	200
RESOLUTIONS	分辨率	10	1073741822

[0116] 进一步的，在步骤S102中，优选的，把森林资源空间数据转化为JSON数据格式，再由JSON数据格式存储为HDFS文件。对于森林资源空间数据而言，一方面这些数据是通过较早的关系型数据库如oracle数据库建立的存储数据，这些数据类型不适用大数据条件下的数据存储和管理，另一方面随着森林资源空间数据的不断扩展，例如包括了更为精确的位置地理信息、特征属性信息等，使得数据的规模容量不断增加，因此有必要将这些已有的森林资源空间数据和不断扩展的森林资源空间数据通过合理的方式转化到适用于大数据环境下数据格式上来。这里，可以通过第三方软件工具，将exce l、csv等格式表示的森林资源空间数据转换为JSON数据格式。而JSON数据格式是适合于在Hadoop的数据集群系统中进行管理的。

[0117] 但是，以JSON数据格式表示的数据通常是一种键值对格式的数据，数据交互的友好性不够。因此，可以借助Hadoop相关的Hive软件建立数据表，然后把JSON数据格式的森林资源空间数据加载到该数据表中，这样就可以利用Hive中类似SQL的HiveQL语言实现数据查询，而所有Hive的数据都存储在Hadoop兼容的文件系统(例如，Amazon S3、HDFS)中。

[0118] 优选的，这里可以是基于Hive软件使用SerDes(序列化器/反序列化器)将结构化

和非结构化的森林资源空间数据转换为JSON数据格式。还可以进一步建立数据表，该数据表把JSON数据格式的森林资源空间数据的名称和地理信息映射为该数据表中的列名和边界形状。

[0119] 优选的，还可以对现有的森林资源空间数据库整体迁移到图2所示的集群系统中，再转换输出为Hadoop对应的HDFS文件。现有的森林资源空间数据库通常是关系型数据库，如oracle数据库，这里可以利用Hive工具软件进行相应的迁移转换。以及，还可以通过一系列的工具包，如Geoprocessing Tools、Esri UDF、Esri Geometry API、Spatial Framework等工具进行转化处理。

[0120] 可见通过步骤S102可以使得森林资源空间数据转换为适于分布式处理的数据集群系统中，能够解决原有森林资源空间数据在数据类型较多、数据量较大情况下统一存储格式和高效存储的问题，避免了数据格式不兼容的问题，并且可以对原有的森林资源空间数据库进行整体迁移，适应大数据的应用需求。并且，经过上述转换为JSON数据格式，以及通过Hive软件建立数据表，具有数据表可扩展的优势，不必受限与原有的关系型数据库中的数据表大小。

[0121] 对于步骤S103，则是基于Hadoop集群的分布式处理特点，在从节点计算机上进行分布式存储。

[0122] 对于步骤S104，建索引步骤：由三个主要阶段组成，即建立分区，构建本地索引和全局索引。

[0123] 对建立分区而言，该阶段将输入文件空间划分为满足三个主要目标的n个分区：(1)块拟合，每个分区应该适合一个尺寸为128MB的HDFS块；(2)空间局部性，空间附近的物体被分配到相同的分区；(3)负载平衡，所有分区的大小应大致相同。为此，通过以下三个步骤加以实现：

[0124] 步骤1：计算分区数。根据分区方程n＝[s(1+α)/B]计算分区数n，其中s是输入文件大小，B是HDFS块容量(如128MB)，α是开销比，默认设置为0.2，这说明了复制记录和存储本地索引的开销。总的来说，这个方程将平均分区大小调整为小于B。

[0125] 步骤2：确定分区边界。在此步骤中，空间数据由最小外包矩形(MBR)进行简化，为了适应具有均匀或偏斜分布的数据，根据构建的基础索引不同，分区边界的计算方式不同。该步骤的输出是表示n个分区的边界的一组n个矩形，它们共同覆盖整个空间域。

[0126] 步骤3：物理分区。给定在步骤2中计算的分区边界，启动MapReduce作业。这里需要决定如何处理可能与多于一个分区重叠的空间范围(例如多边形)的对象。一些索引结构将记录分配给最佳匹配分区，而其他索引结构将记录复制到所有重叠分区。最后，对于分配给分区p的每个记录r，映射函数写入中间对<p,r>。然后，这样的中间对被分组并发送到下一阶段的reduce函数，即本地索引阶段。

[0127] 对构建本地索引而言，在森林资源空间数据集群的从节点计算机上建立本地索引，如R-tree结构的本地索引。此阶段的目的是将所请求的索引结构(例如，Grid或R-tree)构建为每个物理分区的数据内容上的本地索引。这被实现为一个reduce函数，它将分配给每个分区的记录存储在空间索引中，并写入本地索引文件。由于两个原因，每个本地索引必须适合一个HDFS块(128MB)：(1)这允许写入MapReduce程序的空间操作访问在一个map任务中处理每个本地索引。(2)当Hadoop负载平衡器在机器上重新定位块时，它确保本地索引被

视为一个单元。根据第一阶段进行的分区，预计每个分区适合于一个HDFS块。如果分区太大而不能嵌入到一个块中，将其分解成更小的块，每个块大小为64MB，可以写成单个块。为了确保本地索引在连接后保持与块对齐，每个文件都附加虚拟数据(零)，使其完全达到128MB。

[0128] 对构建全局索引而言，则是在主节点上建立一个全局索引，通过该全局索引可以对本地索引进行访问。最终形成了一个包含在从节点计算机12上的本地索引和在主节点计算机11上的全局索引的二级索引机制，由此可以通过索引高效便利的访问所存储的森林资源数据。

[0129] 对于步骤S105，主要是通过工具软件hive和GeometryAPI实现对海量的森林资源空间数据进行统计，并可以把统计结果可视化。该过程实施例如下：

[0130] 优选的，可以先建立统计函数，如下实施例所示：

[0131] create temporary function ST_Bin as'com.esri.Hadoop.hive.ST_Bin';

[0132] create temporary function ST_Point as'com.esri.Hadoop.hive.ST_Point';

[0133] create temporary function ST_BinEnvelope as'com.esri.Hadoop.hive.ST_BinEnvelope';

[0134] 接着，根据业务需求编写HQL(Hive中定义了简单的类SQL查询语言，称为HQL)语句，如下实施例所示：

[0135] FROM(SELECT ST_Bin(0.001,ST_Point(dropoff_longitude,dropoff_latitude))bin_id,*FROM taxi_demo)bins

[0136] SELECT ST_BinEnvelope(0.001,bin_id)shape,

[0137] COUNT(*)count

[0138] GROUP BY bin_id;

[0139] 再将结果存入新建表格，如下实施例所示：

[0140]
```
CREATE TABLE taxi_agg(area BINARY, count DOUBLE)
    ROW FORMAT SERDE 'com.esri.Hadoop.hive.serde.JsonSerde'
    STORED AS INPUTFORMAT 'com.esri.json.Hadoop.UnenclosedJsonInputFormat'
    OUTPUTFORMAT 'org.apache.Hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat';
```

[0141] 最后，将统计结果进行可视化显示，如通过可视化工具WebGIS进行显示。

[0142] 通过上述方式，本发明实施例公开一种森林资源空间数据的分布式管理方法，该方法包括步骤构建集群系统、数据转换、数据存储、建立索引和统计调用。通过该方法能够把多种类型的森林资源空间数据统一转化为适用于Hadoop集群系统的HDFS文件进行存储，并通过hive、spatialHadoop等软件工具实现对森林资源空间数据的高效统计，具有兼容性强、效率高以及技术开发成本低等优势。

[0143] 以上所述仅为本发明的实施例，并非因此限制本发明的专利范围，凡是利用本发明说明书及附图内容所作的等效变换，或直接或间接运用在其他相关的技术领域，均同理包括在本发明的专利保护范围内。

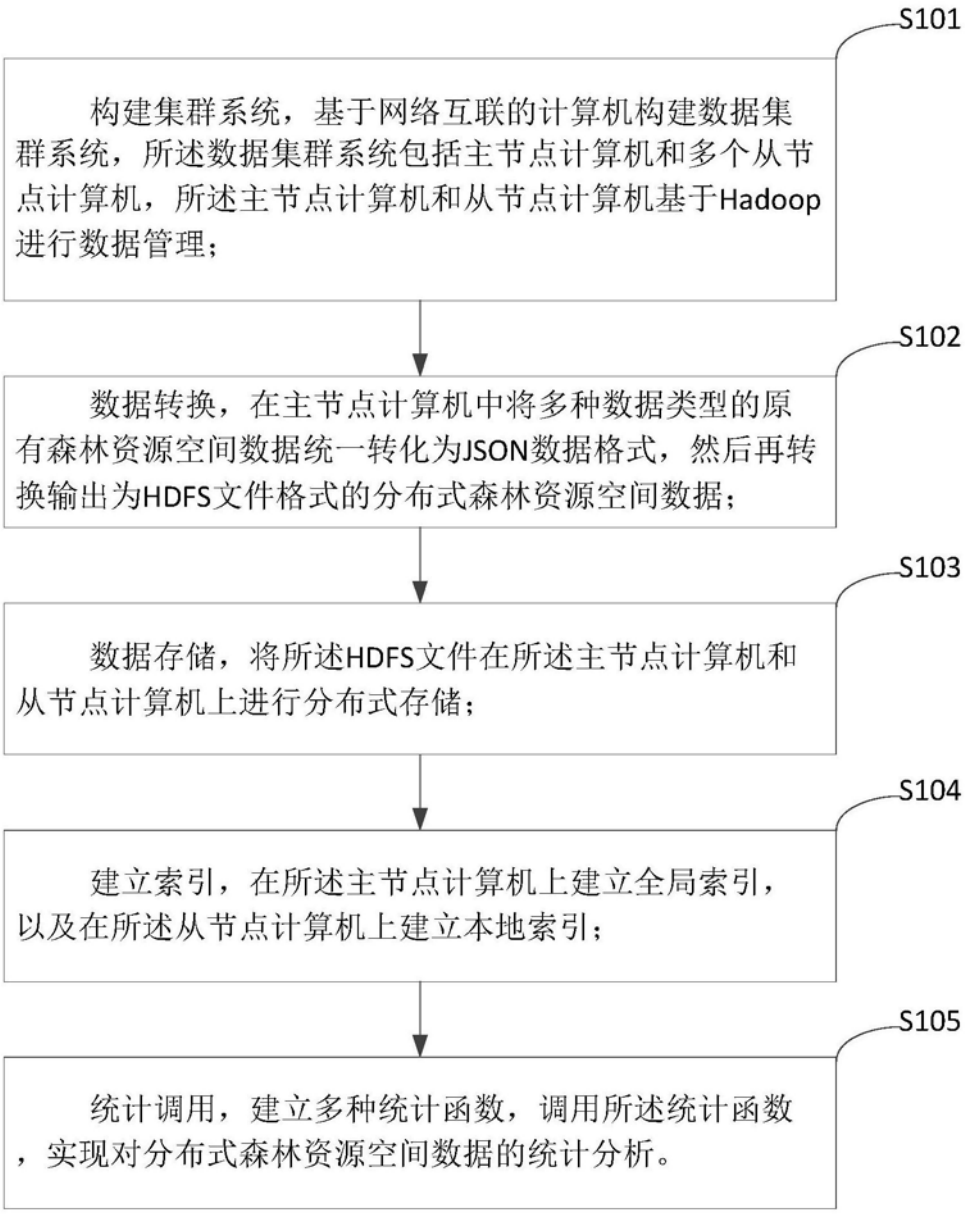

S101
构建集群系统，基于网络互联的计算机构建数据集群系统，所述数据集群系统包括主节点计算机和多个从节点计算机，所述主节点计算机和从节点计算机基于Hadoop进行数据管理；
S102
数据转换，在主节点计算机中将多种数据类型的原有森林资源空间数据统一转化为JSON数据格式，然后再转换输出为HDFS文件格式的分布式森林资源空间数据；
S103
数据存储，将所述HDFS文件在所述主节点计算机和从节点计算机上进行分布式存储；
S104
建立索引，在所述主节点计算机上建立全局索引，以及在所述从节点计算机上建立本地索引；
S105
统计调用，建立多种统计函数，调用所述统计函数，实现对分布式森林资源空间数据的统计分析。

图1

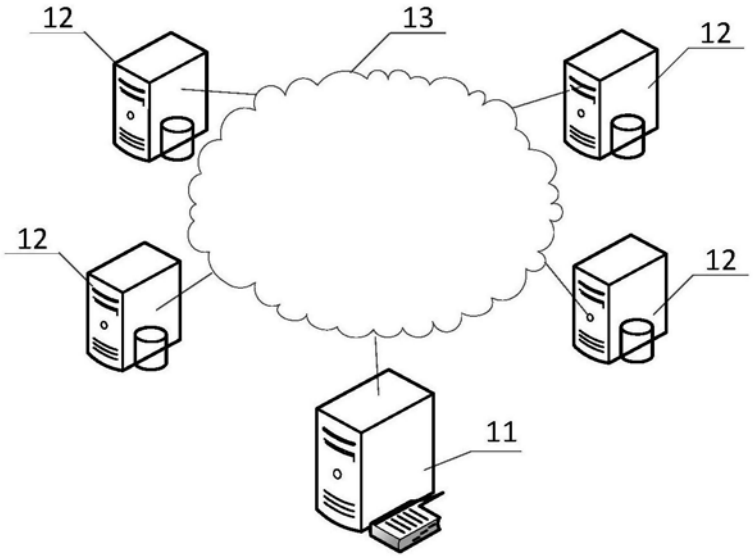

12
13
12
12
12
11

图2

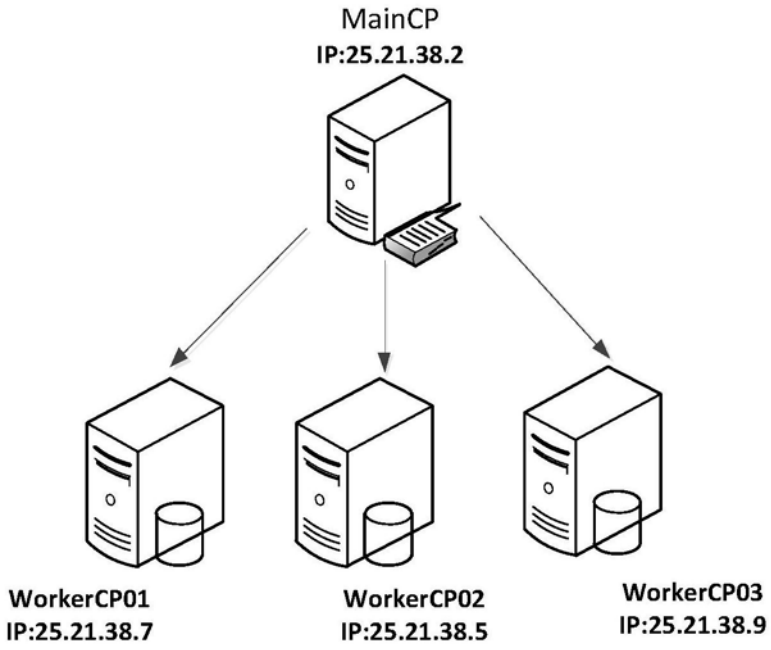

MainCP
IP:25.21.38.2
WorkerCP01
IP:25.21.38.7
WorkerCP02
IP:25.21.38.5
WorkerCP03
IP:25.21.38.9

图3

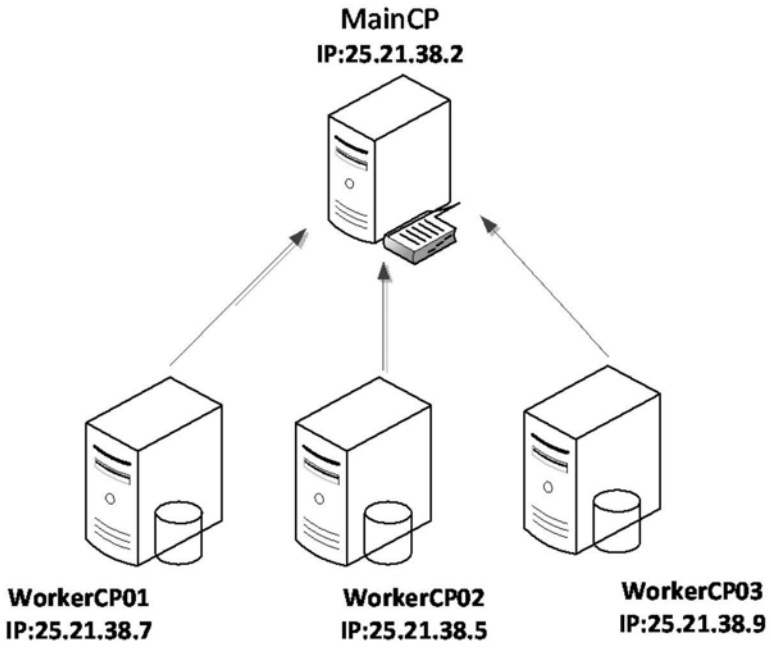

MainCP
IP:25.21.38.2
WorkerCP01
IP:25.21.38.7
WorkerCP02
IP:25.21.38.5
WorkerCP03
IP:25.21.38.9

图4

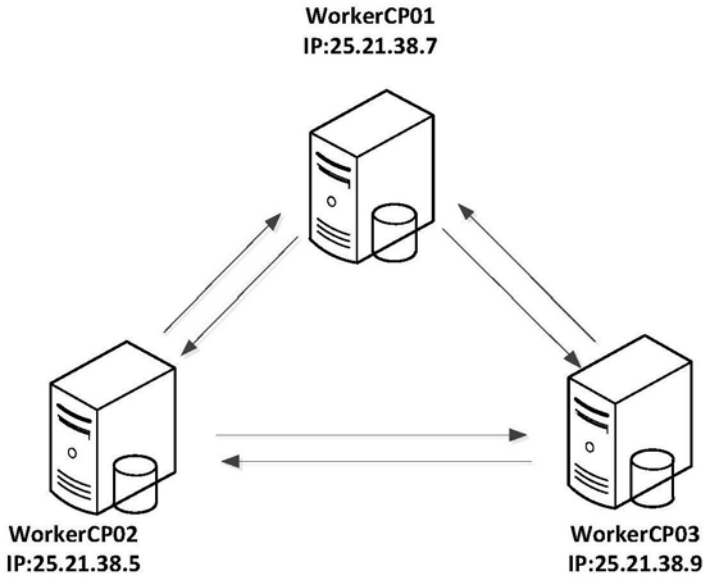

WorkerCP01
IP:25.21.38.7
WorkerCP02
IP:25.21.38.5
WorkerCP03
IP:25.21.38.9

图5

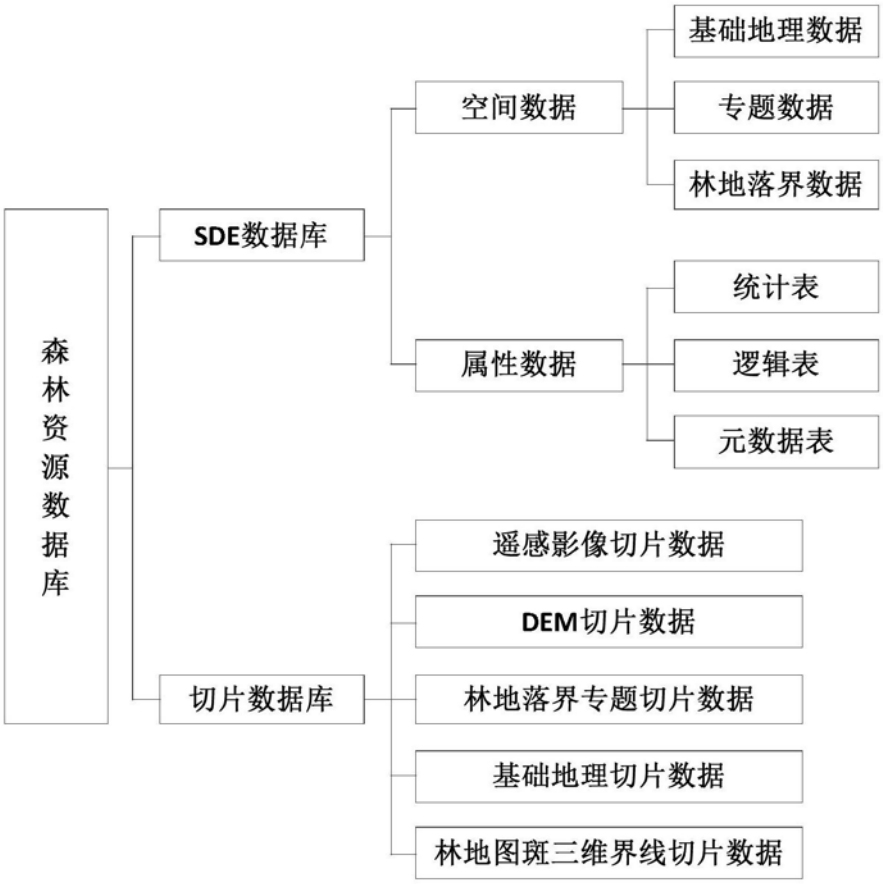

森林资源数据库
SDE数据库
空间数据
基础地理数据
专题数据
林地落界数据
属性数据
统计表
逻辑表
元数据表
切片数据库
遥感影像切片数据
DEM切片数据
林地落界专题切片数据
基础地理切片数据
林地图斑三维界线切片数据

图6