白铁男,唐维尧,金石声,等. 数据湖技术在气象卫星资料业务共享中的应用[J]. 中低纬山地气象,2023,47(1):112 - 115.

数据湖技术在气象卫星资料业务共享中的应用

白铁男,唐维尧,金石声,谭海波

(贵州省气象信息中心,贵州　贵阳　550002)

摘　要:文章依托大数据云平台“天擎”系统的共享网络存储(Network Attached Storage,以下简称“NAS”),通过省级气象通信系统的文件分发策略,将气象卫星广播数据迁移到共享 NAS 上进行统一存储。基于 B/S 系统构架在贵州省搭建气象数据湖集群,实现对“天擎”系统共享 NAS 存储权限的统一管理。数据湖提供了前端图形界面对目录权限和账户权限统一管理,一方面规范了数据访问权限,保障数据和账户的安全性;另一方面解决了因数据反复迁移、多处存储而造成的资源浪费,符合气象业务的发展需求。

关键词:天擎;数据湖集群;CMACast;NAS 存储;安全管理

中图分类号:TP311.56　**文献标识码**:B

The Application of Data Lake Technology in The Sharing of Meteorological Satellite Data Service

BAI Tienan,TANG Weiyao,JIN Shisheng,TAN Haibo

(Guizhou Meteorological Information center,Guiyang 550002,China)

Abstract:Relying on the shared NAS storage of the big data cloud platform, Tianqing system, and through the file distribution strategy of the provincial meteorological communication system, the meteorological satellite broadcast data is migrated to the shared NAS for unified storage. Based on the B/S system architecture, a meteorological data lake is built to uniformly manage the data on the shared NAS storage of the big data cloud platform Tianqing system. Data Lake provides a front - end graphical interface for unified management of directory permissions and account permissions. On the one hand, it regulates data access permissions and ensures the security of data and accounts. On the other hand, it solves the waste of resources caused by repeated data migration and multiple storage, which meets the development needs of meteorological services.

Key words:Tianqing;Data lake cluster;CMACast;NAS storage;security management

0　引言

随着大数据时代的到来,气象业务对数据的需求量呈现指数级增长[1]。常规地面和高空等气象数据已不能满足现阶段气象预报的需求[7],数值预报等卫星广播气象数据如何存储和共享成了气象工作者广泛关注的问题[2]。中国气象局卫星数据广播系统(以下简称“CMACast”)自 2012 年建成以来,大幅增加了广播气象数据的种类和下发量[3]。作为贵州省接收广播气象数据的主要渠道,系统采用卫星天线小站接收数据的方式,其程序部署在省级国内气象通信系统(以下简称“CTS”)的服务器上,在完成数据下载后通过 FTP 传输到共享服务器进行存储,并向业务单位提供 SAMBA、FTP 等共享访问服务[4]。由于各共享服务器独立管理,数据的存储目录和用户的访问权限管理混乱,缺乏统一管理的有效手段[6]。为满足气象业务发展的需求,解决上述问题,依托大数据云平台“天擎”系统(以下

收稿日期:2021 - 11 - 08

第一作者简介:白铁男(1991—),男,工程师,主要从事气象信息技术实时业务监控工作,E - mail:Z1114257065@foxmail.com。

简称“天擎”)在国省协同部署的背景,在“天擎”公共 NAS 存储上规划目录结构,搭建数据湖集群,在数据湖的前端目录视图上实现对目录和账户权限的统一管理。

1　流程构架

“数据湖”是通过将原始数据分类存储到不同数据池,并在各数据池里将数据整合转化成容易分析的统一存储格式进行存储,以方便用户对大量原始数据池中的数据加以分析利用,从而产生经济效益的新兴技术[5]。“天擎”系统作为提供各个业务部门使用气象业务资料的数据环境,所辖的 NAS 是“天擎”的组成部分,用来存储原始资料,或者是加工的中间产品和最终产品。数据湖集群提供数据共享服务需要挂载“天擎”NAS,为避免因反复挂载而造成资源浪费和安全性问题,将数据湖服务直接部署在“天擎”的子模块产品加工系统(DPL)的 4 个节点上,其中 2 个作为数据湖的管理节点(Master),另外 2 个作为计算节点(Worker),以主备模式运行,主节点停止服务时自动切换到备份节点上工作。ansible 作为自动化运维工具,基于 Python 开发,其集合众多运维工具的优点,不需要安装代理或者客户端且简单易上手。通过 ansible 自动化运维工具实现数据湖集群的系统配置、程序部署和运维管理。需要注意 ansible 自动化工具必须和 PostgreSQL 数据库的主节点在同一台服务节点上部署。数据湖软件的部署环境为 Redhat 7.5,涉及的系统软件主要包括 Java 开发工具包 JDK、系统元数据库 PostgreSQL、系统缓存数据库 Redis、搜索服务工具 ElasticSearch 以及分析和可视化平台 Kibana。ElasticSearch 主要负责收集文件的下载信息,Kibana 负责将 ElasticSearch 获取的监控信息在前端界面展示,以“运维监控”模块集成到数据湖的前端界面,实现对数据资源下载情况的实时监视,数据下载监视界面如图 1 所示。

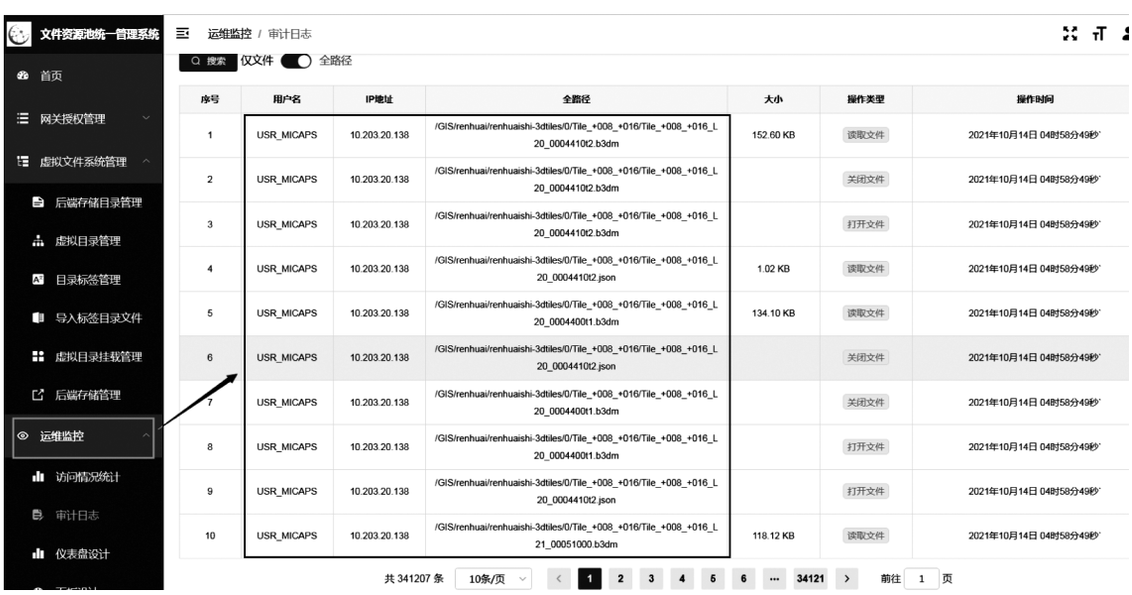

图 1　数据下载监视界面

Fig. 1　The monitoring interface of Data download

涉及的应用软件包括数据管理服务 fsgateway - rest 和数据存储服务 fsgateway - gw。fsgateway - rest 部署在管理节点上,fsgateway - gw 部署在计算节点上。fsgateway - rest 作为前端管理界面进程,将前端界面的目录权限配置写入元数据库,fsgateway - gw 作为工作节点,会根据元数据库将虚拟目录的访问权限赋予相应的用户。具体工作流程构架如图 2 所示。数据湖除了服务集群还包括众多的的客户端,数据湖的客户端是提供给 Linux 和 Windows 系统连接数据湖文件管理系统的 1 个服务,它类似于 1 个 NAS,可以把目录信息直接挂载到系统的某个目录上,用户通过安装客户端,根据需要访问不同的目录。系统管理员可以通过可视化前端界面对用户的使用权限统一进行管理。

2　部署方法

数据湖部署安装过程分为服务端部署和客户端部署,服务端安装部署过程包括上传安装部署包、安装 ansible 自动化部署软件、配置服务器登录信息和分组、配置软件部署信息、执行安装命令等。

整个部署过程需要使用 root 用户进行，首先上传 fsgateway – manager – 1.0.0.tar.gz 安装部署包到管理节点，通过 tar 命令解压压缩包，再通过 yum 命令安装 ansible 自动化部署软件，ansible 软件下的 Playbooks 批量安装命令可对多个节点上的软件同时进行部署安装，后面对软件的部署以及检查都会用到。分别修改 inventory 和 common.yml 配置文件完成软件的分组信息和部署信息配置，最后执行 ansible – playbook – i inventory site.yml 完成批量安装。

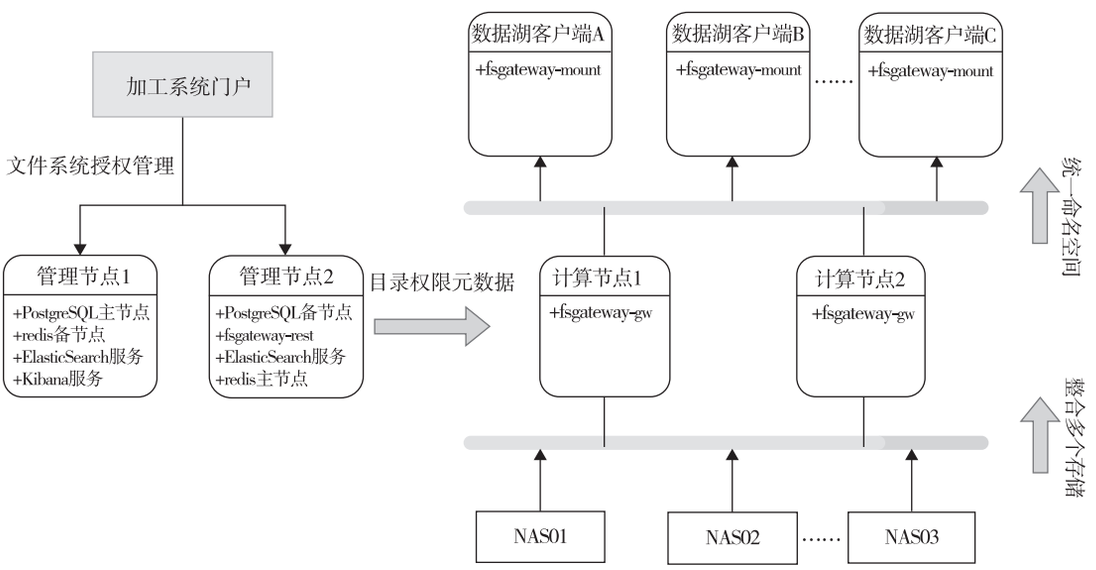

图 2　数据湖流程构架

Fig. 2　The process architecture of data lake

在 Linux 和 Windows 系统环境下，数据湖客户端需要分别进行部署。Linux 系统环境的客户端需要依赖用户态 fuse 程序，安装时需要先将安装包 fsgateway – mount.tar.gz 上传至客户端服务器实现解压，再到 conf 目录下修改 fs – env.sh 配置文件。配置文件的参数 GW_IP 修改为本省数据湖计算节点地址，参数 source 修改为通过数据湖前端挂载的虚拟路径，参数 target 修改为本机创建的访问路径。Windows 系统环境下安装客户端要求安装路径不能有中文，解压压缩包后双击 winfsp – 1.8.20304.msi 程序弹出安装导航，后续全部选择默认选项即可。再到 config 目录下修改 cmafs.ini 配置文件，参数 serverip 修改为本省数据湖计算节点地址，双击 MUSIC_dataDirSer.exe 快捷方式弹出文件访问客户端界面，由于数据湖可以直接读取“天擎”业务用户库表信息，将“天擎”的账户名和密码直接输入即可。Linux 环境的客户端需要绑定服务器 IP 地址将目录的访问权限赋予给系统用户，而 Windows 环境的客户端是通过绑定“天擎”的业务账户将目录的访问权限赋予给用户使用，无需提供 IP 地址，文件访问 Windows 客户端情况如图 3 所示。

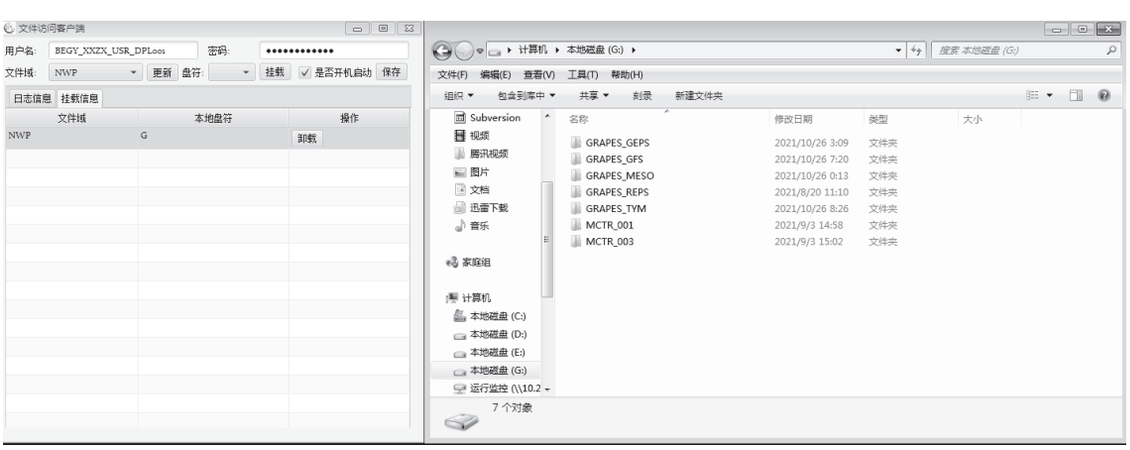

图 3　Windows 文件访问客户端

Fig. 3　The client of windows file access

3　资料共享实现方法

CMACast 程序部署在 CTS 集群的 5 号机和 6 号机上,通过 Veritas Cluster Server(简称“VCS”)集群管理软件上的 dvb - grp 程序进行统一调度和管理。CMACast 从不同的通道接收并下载气象资料后通过 FileTrans - grp 程序对各通道下的资料 FTP 传输到“天擎”共享 NAS 存储规划的目录下进行统一保存。再通过 CTS 的前端界面新增 SJH_GX 分发用户,根据规划的目录结构将以往推送共享服务器的 CMACast 资料推送“天擎”共享 NAS 下进行存储,配置界面如图 4 所示。

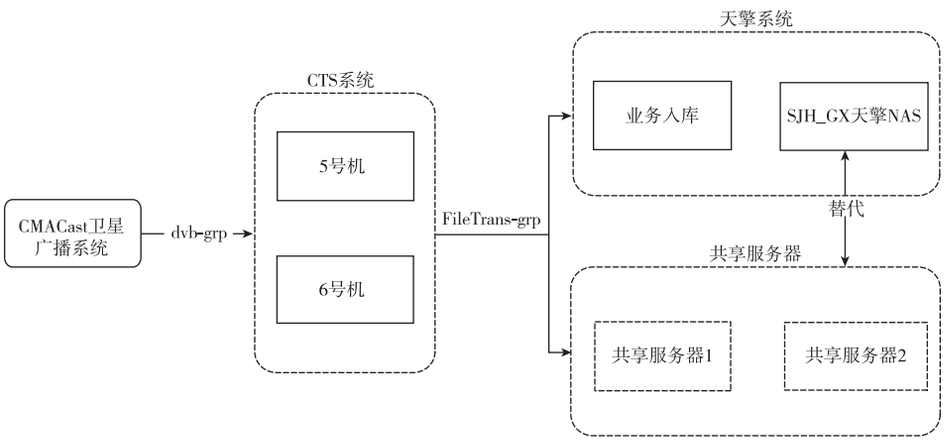

图 4　CTS 文件推送策略界面

Fig. 4　The strategy interface of CTS file push

通过数据湖前端界面可以实现对“天擎”NAS 存储路径的管理。首先将 CMACast 资料存储目录结构导入数据湖,通过命令 find /CMADAAS/DATA/SHARE/cmacast - type d > > cmacast. csv 将 NAS 的存储目录结构递归到 cmacast. csv 文件,再通过数据湖前端界面的虚拟文件系统管理—导入标签目录文件模块将 NAS 的目录结构导入数据湖。需要注意导入模板的第 1 列是 NAS 存储的真实目录,第 2 列是虚拟目录可随意编写,通过客户端看到的目录就是虚拟目录。数据湖的用户分为普通用户、操作系统用户和认证用户。普通用户即“天擎”的业务用户,操作系统用户是指为融入业务系统而创建的用户,认证用户是固定的 lakefsuser。通过配置文件 fs - env. sh 在 Linux 客户端实现目录挂载,认证用户与客户端的 IP 地址绑定,同时客户端的 IP 地址与普通用户或操作系统用户绑定,这样只需要通过授权管理模块将 NAS 目录的访问权限授予普通用户或操作系统用户,即能在客户端实现对已授予权限目录下的资料进行访问,通过界面也可以设置可读写权限以及子层目录的访问权限。

4　结束语

现阶段气象业务对数据的安全和监管要求越来越高,因此亟需安全且可靠的系统来对这些海量级的气象资料统一进行管理。本文依托“天擎”产品加工系统的硬件基础,通过部署数据湖集群,实现对 CMACast 卫星广播资料统一推送存储和权限管理。一方面,统一在“天擎”NAS 进行存储,避免了数据因反复迁移而造成的资源浪费,取代部分共享服务器。另一方面,通过数据湖前端界面导入的 NAS 目录对应 1 个虚拟目录,客户端只能看到虚拟目录,在一定程度上提高了数据的安全性。最后,数据湖的 PostgreSQL 同步“天擎”系统的账户数据库,使用“天擎”系统的账户即可实现数据访问,避免重复申请账号的问题,更加方便运维管理,符合当前对数据安全和账户安全管理的要求。

参考文献

[1] 马驰. 民机运行支持数据湖设计与实现[J]. 计算机测量与控制,2021(29):175 - 179.

[2] 刘红艳,张晓庆,吕晓丹. 气象工作中 CMACast 系统数据资料的备份存储分析[J]. 吉林农业,2015(6):119 - 120.

[3] 韩超. 基于 SSH 协议的 CMACast 智能维护系统设计[J]. 电脑知识与技术,2019(31):68 - 71.

[4] 杨鹤,马云飞,李泽,等. 浅述 CMACast 系统在 Linux 中的安装配置[J]. 科学与信息化,2018(10):30.

[5] 郭文惠. 数据湖——一种更好的大数据存储构架[J]. 电脑知识与技术,2016(30):4 - 6.

[6] 李珏,谭海波,李波,等. 气象虚拟化云平台搭建及安全防护的探讨[J]. 中低纬山地气象,2021,45(4):108 - 110.

[7] 谭海波,汪华,金石生,等. 气象区域站数据全流程监控系统的设计和实现[J]. 中低纬山地气象,2021,45(6):111 - 115.