何林,吉庆. 气象-环保数据交换共享方案的设计与实现[J]. 陕西气象,2017(1):36-39.
文章编号:1006-4354(2017)01-36-04

气象-环保数据交换共享方案的设计与实现

何 林[1],吉 庆[2]
(1. 陕西省气象信息中心,西安 710014;2. 渭南市气象局,陕西渭南 714000)

摘 要:为了有效解决气象部门和环保部门实时监测数据交换共享的问题,提出了基于专网的数据库同步、基于公网 FTP 的数据交互、基于网络爬虫的数据抓取三种共享方案,并对三种方案的实现方法进行了论述,通过对各个方案优缺点的对比,给出了不同应用场景下的最优共享策略。

关键词:气象-环保数据;共享方案;数据库同步;FTP;网络爬虫
中图分类号:P409 **文献标识码**:B

数据挖掘、大数据分析等技术在各个行业逐渐由基础研究转向实际业务应用,某个行业的数据分析往往需要以其它行业的数据为支撑,因此不同行业或部门间的业务数据交互越来越多。以气象部门和环保部门为例,在酸雨、沙尘暴、雾霾等天气越来越被关注的背景下,气象部门与环保部门需要打破数据壁垒,结合各自的专业优势在环境气象监测、预报预警等方面展开研究与合作。由于气象部门和环保部门分属不同的局域网,且双方的数据传输、监测、存储、内部共享、应用方式等都不尽相同。因此,需要根据双方的应用特点,制定一个安全合理的底层数据共享与交换方案,为监测数据的业务应用奠定基础。

1 气象-环保跨行业监测数据的特点

气象部门和环保部门的底层数据主要为通过各类采集仪器收集到的监测数据,基于通用分组无线服务技术(General Packet Radio Service,下简称"GPRS")或地面有线网络,以文件或数据流的方式进行自下而上的分级传输和集中收集,解析后多以结构化数据形式存储于数据库管理系统中,用以开展业务研究。常用的环境监测类型包括环境空气、降水、地表水、饮用水、地下水、邻近海域水质、土壤、噪声、废气、废水、生物监测等 11 种[1]。而气象观测则包括地基(自动气象站、酸雨观测站、大气成分站、海上漂浮站、太阳辐射站等)、空基(探空气球、多普勒雷达、探空飞机、无人机)、天基(极轨卫星、静止卫星、空间站)等三大类型。管理这些监测数据的部门一般为国家级、省级、地市级和县级的环境监测总站及气象部门。

气象-环保监测数据种类繁多、结构复杂,且

收稿日期:2016-09-22
作者简介:何林(1987—),男,汉族,陕西武功人,硕士,工程师,从事气象信息软件研发、气象数据集约化环境建设。
基金项目:陕西省气象局研究型业务重点科研项目(2015Z-6)

[10] 肖晶晶,霍治国,李娜,等. 小麦赤霉病气象环境成因研究进展[J]. 自然灾害学报,2011,20(2):146-152.
[11] 徐崇浩,何险峰,刘富明,等. 四川小麦赤霉病流行的气象条件及其时空分布规律和大气环流背景[J]. 西南农业学报,1996,9(3):60-67.
[12] 张吉昌,王捍东,白庄君,等. 汉中盆地小麦赤霉病发生规律及防治技术[J]. 陕西农业科学,1995,(2):24-25.
[13] 曹祥康,陈爱光,田平阳. 福建省小麦赤霉病气候预报初探[J]. 中国农业气象,1994,15(3):33-35+27.
[14] 左豫虎,郑莲枝,张匀华,等. 黑龙江省春小麦赤霉病流行的预测方法[J]. 植物保护学报,1995,22(4):297-302.
[15] 许昌燊. 农业气象指标大全[M]. 北京:气象出版社,2004:164.
[16] 商鸿生. 麦类作物病虫害诊断与防治原色图谱[M]. 北京:金盾出版社,2003:44-45.

存在一定的交叉,对这些监测数据进行共享大体可分为实时与历史数据的共享[2]。对于历史数据,在双方同意且配合的前提下,直接使用从源数据库中导出、导入、归档的方式实现行业间的数据交换;而对于实时数据,则需要在解决网络连通性的前提下,通过数据库同步或编写程序实现共享。本文主要在跨行业的实时数据共享交换方面进行分析研究。在保证数据安全的前提下,提出基于专线网络的数据库同步、基于公网 FTP 的数据交互、基于网络爬虫技术的三种数据共享方案。

2　基于专线网络的数据库同步

2.1　基本原理

搭建专线网络,保证气象部门与环保部门内部局域网的连通性,编写触发器脚本或借助数据库管理系统(Database Management System,简称 DBMS)自带的同步功能[3],实现数据交换。

2.2　设计流程

可进行数据库同步的前提是网络连通性,本方案直接使用专线网络实现两个部门的网络连通,同时需对防火墙进行配置,开放数据库软件的相关端口[4],此时,远端数据库就相当于本地库可以进行相关的操作。以气象部门单向同步环保部门数据为例(图 1),即环保部门为数据提供者,数据已实时存储于 TAB_SOUR 表中;气象部门为数据接收者,数据将存放于 TAB_DEST 表中。同步数据时,首先须在目标库中建立数据库对象,包括 TAB_SOUR 的表结构及相关索引、视图、存储过程等,从源库中导出它们的创建脚本,并在目标库中执行;其次建立数据库链接 DBLINK_A&B,实现跨库数据访问[5];最后就是编写触发器、同步脚本,或使用数据库本身自带的同步工具(如主流的 Oracle 的 DataGuard、Streams 方式、SQL SERVER 的 SQLJOB 及发布/订阅方式、MySql 的主从复制功能等),实现单向的数据交换。环保部门同步气象数据的实现流程类似于气象部门单向同步环保部门数据的流程。

图 1　基于专线网络的数据库单向同步流程

2.3　适用场景

本方案适用于较为正式的气象部门与环保部门数据合作,租用专线网络时需要一定的成本,网络带宽和稳定性将对数据同步的效果产生决定性的影响。另外,当双方 DBMS 类型一致时,数据同步相对容易,若类型不一致,则需借助其它软件进行数据同步,从而增加了研发和维护成本。

3　基于公网 FTP 的数据交互

3.1　基本原理

借助气象部门或环保部门任一方已有的互联网资源,搭建公网 FTP,以此为桥梁,将需共享的数据以文本形式通过 FTP 进行传输,双方只需通过 PUT/GET 进行上传/下载并导出/导入数据即可。

3.2　设计流程

本方案通过公网 FTP 间接实现网络连通,数据同步的关键是编写代理软件[6]。以气象部门单向同步环保部门数据为例(图 2),首先需在气象部门的数据库中导入环保数据库中相关表结构、索引、视图等对象。其次在气象部门的内网服务器上运行代理进程 Agent_Master,在环保部门的内网服务器上运行代理进程 Agent_Slave。环保部门同步气象数据的实现流程类似于气象部门单向同步环保部门数据的流程。

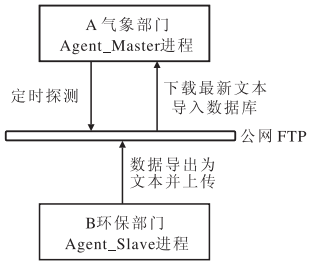

图 2　基于公网 FTP 的数据交互流程

Agent_Slave 进程实现伪代码:

〈using 引用源部门数据库平台的名字空间〉

function_Slave

{

string dbstr="源数据库服务器 IP,数据库名称,用户名,密码等";

使用 dbstr 创建一个数据库连接对象 conn;

conn. open();打开连接
string sqlsour="select * from TAB_SOUR where insertime between … and …";
//查询源表两次监测间隔时间之间的数据
获取检索结果并存入数组 data[][]
while(遍历检索结果)
{
string sqldest="insert into TAB_DEST values(data[][0],data[][1]…)";
//构造数据表插入语句
追加写入文本文件 sql. txt;
}
conn. close();关闭连接
string ftpser="FTP 服务器 IP,用户名,密码,传输模式等";
ftpser. connect();打开 FTP 连接
ftpser. put(sql. txt);上传文件
ftpser. close();关闭 FTP 连接
}

Agent_Master 进程实现伪代码:

〈using 引用目标部门数据库平台的名字空间〉
function_Slave
{
string ftpser="FTP 服务器 IP,用户名,密码,传输模式等";
ftpser. connect();打开 FTP 连接
if(存在最新上传的 sql. txt 文本)
ftpser. get(sql. txt);下载文件
ftpser. close();关闭 FTP 连接
string dbstr="目标数据库服务器 IP,数据库名称,用户名,密码等";
使用 dbstr 创建一个数据库连接对象 conn;
conn. open();打开连接
While(逐行遍历 sql. txt)
conn. execute(逐行执行插入语句)
conn. close();关闭连接
}

3.3 适用场景

本方案适用于资料种类单一、不是特别正式的气象-环保数据共享。相比专线网络的方案,在实现网络连通方面直接借助公网资源,节约了成本。但由于需在双方服务器上部署代理程序,尤其是数据源方完全占据主动权,一旦停止 slave 代理程序,相当于数据源丢失,通过 FTP 为媒介的数据传输就会中断。因此双方必须在数据同步前达成共识,对代理程序进行必要的监控与维护,才能保证数据的持续、稳定交换。

4 基于网络爬虫技术的数据抓取

4.1 基本原理

网络爬虫,又称网页蜘蛛,是指按照一定的规则,自动地抓取特定网站信息的程序或者脚本。由于气象部门及环保部门数据一般都会通过 WEB 页面的形式在其官网进行开放性的展示,因此可利用爬虫技术,对感兴趣的在线数据进行抓取并存放于本地数据库中,从而实现单向的数据交换。如果网站已经提供了应用编程接口(Application Programming Interface,简称 API),以标准化、结构化的格式来共享数据,则可直接通过 API 下载数据。然而,目前大多数网站会将维护前端界面比后端 API 置于更高的优先级,几乎不提供和维护 API 数据接口,因此爬虫就显得必不可少了[7]2-7。爬虫技术不受共享双方后台数据库架构、数据类型、表结构的限制,数据抓取较为灵活。

4.2 设计流程

爬虫脚本可通过多种编程语言实现。以 Python 语言为例,Python 是一种面向对象、解释型的计算机程序设计语言,语法清晰简洁、易读易维护,已成为一种广受好评的流行性编程语言。在实现网页抓取时,Python 提供了正则表达式、Beautiful Soup 模块、Lxml 模块三种方式[7]26-31,虽然这三种抓取方法在性能、难度方面各有差异,但基本实现流程是类似的。图 3 给出了使用正则表达式法进行网页数据抓取,并将感兴趣的数据存入 Oracle 数据库的基本流程。

数据抓取需完成相应的准备工作。在数据库方面,需要设计存储数据的表结构、编写并执行脚本完成数据表的创建;在待抓取数据的网站方面,需要检查 robots. txt 文件以明确页面是否允许被抓取数据,检查 Sitemap 网站地图以定位页面层级,检查网站构建技术类型,估算页面数据量等。

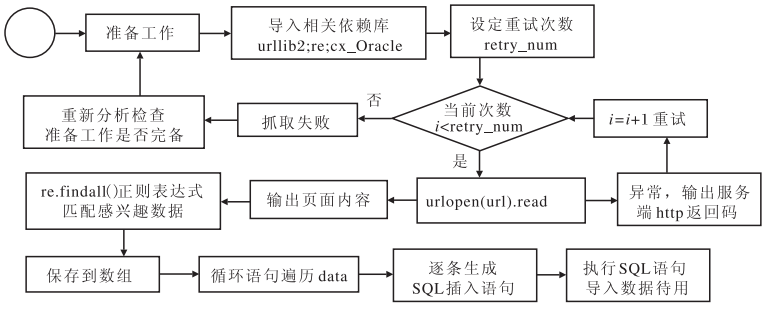

图 3　基于网络爬虫抓取数据并进行存储的流程

这些准备工作是否完备往往决定了爬虫程序能否成功地抓取目标页面的数据。

4.3　适用场景

本方案主要针对气象或环保部门公开数据的共享,且被共享的数据需具备网页展示的载体,是一种非正式的数据共享。如果被共享方的网站页面设计较为复杂,或制定了限制网络爬虫下载频率、速度、封禁 IP 等策略,则会对数据共享的持续性、稳定性造成一定影响。此外,气象、环保的历史数据不一定全部对外开放下载,如果涉及到具有密级的关键数据,最好是通过对方允许才可抓取,否则可能引起侵犯版权等法律问题。

5　三种方案的对比

三种气象-环保跨行业数据交换的方式各有优缺点(表 1),在实际应用时,应综合考虑各方因素制定最优策略。

表 1　三种气象-环保数据交换共享方案对比

对比项目	基于专线网络	基于公网 FTP	基于网络爬虫
同步频次	实时同步	定时同步	定时同步
网络构建成本	高	中	低
研发运维成本	低	中	高
单次数据量	小	较大(取决于同步频次)	较大(取决于同步频次)
资料种类相关性	低	较高	高
数据库相关性的影响	双方一致最佳	影响不大	无影响
性能瓶颈	网络带宽及稳定性	代理程序的正常运行	待抓取页面的限制
使用场合	正式	半正式	非正式

总的来说,如果是较为正式的数据合作,需共享的资料种类较多,且有一定资金支持,最好选择基于专线网络的共享方案。如果仅需共享个别资料,合作不是特别正式,从投入成本方面考虑,则可对各类资料的特点进行逐一分析,研发代理程序或爬虫程序,选择基于公网 FTP 和基于网络爬虫技术的方案进行数据交换,则相对较为合理。

参考文献:

[1]　HJ 660—2013 环境监测信息传输技术规定[S]. 北京:中国环境科学出版社,2013.

[2]　何林,范涛,曹波. 区域自动气象站数据库整合设计与实现[J]. 陕西气象,2014(4):44-46.

[3]　潘承斌. 利用触发器实现数据同步[J]. 电脑编程技巧与维护,2010(10):60-62.

[4]　陈艺宏,林荣惠,张磊. VPN 技术在气象网络中的应用[J]. 陕西气象,2010(5):37-38.

[5]　何伟,郝雅青,周利斌. 基于网络隔离的数据库同步方法[J]. 信息安全与通信保密,2010(1):82-84.

[6]　赵洁. 同步 FTP 上载/下载程序的实现技术[J]. 计算机系统应用,2002(6):38-40.

[7]　Lawson Richard. 用 Python 写网络爬虫[M]. 李斌,译. 北京:人民邮电出版社,2016.