**亚信联创流量经营2.01**

**SCA产品之数据采集模块**

**项目说明文档**

（v0.1）

亚信联创科技（中国）有限公司

亚信联创科技（中国）有限公司

亚信联创科技（中国）有限公司

二零一三年七月

**文档修改记录**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **序号** | **日期** | **版本** | **修改内容说明** | **修改人** | **审阅人** | **审阅日期** |
| 1 | 2013.7.28 | 1.0 | 创建文档 | 潘国远 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

目录

[1 功能概述 4](#_Toc360001225)

[2 运行环境示意图 4](#_Toc360001226)

[3 功能详细说明 4](#_Toc360001227)

[3.1 配置文件 4](#_Toc360001228)

[3.2 源数据说明 6](#_Toc360001229)

[3.3 数据处理主体流程： 7](#_Toc360001230)

[3.4 其他细节处理逻辑 9](#_Toc360001231)

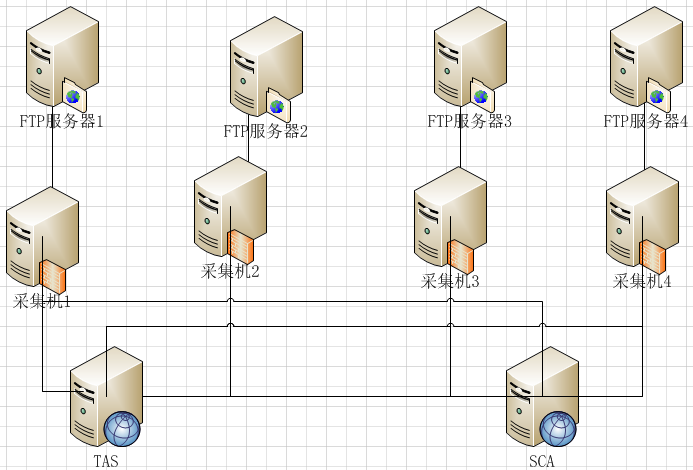
[3.5 根据测试进度待补充 10](#_Toc360001232)

**数据采集模块项目说明**

# 功能概述

预处理模块的主要功能是采集FTP服务器上的日志文件到采集机上，然后做分类、合并输出到TAS和SCA服务器上。

# 运行环境示意图



# 功能详细说明

## 配置文件

模块的正确运行需要根据具体运行环境配置3类参数：

1. 分别是采集数据的源路径
2. 数据采集分类后需要转移到的源服务器上的备份路径
3. 数据采集分类后需要转移到的远程HADOOP集群上的HDFS的路径

***(a)预处理文件采集的配置（参照：fetch.xml）***

<property key=*"fetch.module.switch"* desc=*"数据采集开关"* type=*"string"*>on</property>

<property key=*"scan.time.interval.value"* desc=*"采集线程的扫描周期 "* type=*"int"*>1</property>

<property key=*"scan.ago.folder"* desc=*"扫描多长时间以前的文件夹(单位：分钟) "* type=*"int"*>15</property>

<property key=*"scan.time.interval.unit"* desc=*"扫描时间间隔单位"* type=*"string"*>minute</property>

<property key=*"tmp.thread.start.interval.time"* desc=*"扫描临时目录线程启动时间间隔,单位是秒，注意此选项要比扫描诺西目录的时间间隔必须要短"* type=*"int"*>20</property>

<property key=*"remote.base.path"* desc=*"诺西ftp路径"* type=*"string"*>E:/export/home/sca/pretreatment/remotepath</property>

<!-- <property key="remote.base.path" type="string">ftp://ftpuser:12345@192.168.253.128/var/ftp/data</property> -->

<property key=*"native.base.path"* desc=*"本地路径"* type=*"string"*>E:/export/home/sca/pretreatment/data</property>

<property key=*"native.tmp.base.path"* desc=*"临时目录"* type=*"string"*>E:/export/home/sca/pretreatment/download-tmp</property>

<property key=*"invalidate.base.path"* desc=*"存放诺西异常文件目录"* type=*"string"*>E:/export/home/sca/pretreatment/invalidate</property>

<property key=*"fetch.info.base.path"* desc=*"文件夹记录文件目录 "* type=*"string"*>E:/eclipse/workspace/fetchinfo/</property>

<property key=*"whole.filename.regex"* desc=*"扫描文件名正则表达式"* type=*"string"*>\d{2}-\d{2}-\w+-\w+-\d{2}\_\d{14}\_\d{2}~{0,1}</property>

<property key=*"file.transfer.threads"* desc=*"文件并行下载线程数"* type=*"int"*>50</property>

<property key=*"scan.folder.regex"* desc=*"扫描文件夹正则表达式"* type=*"string"*>^(\d{12})$</property>

<property key=*"tmp.folder.regex"* desc=*"临时文件夹表达式"* type=*"string"*>^(\d{12})$</property>

<property key=*"scan.date.interval"* desc=*"扫描文件夹记录文件数,在配置此选项时请注意,比如诺西存放的文件是3天，那么里的参数必须大于等于3天，要不然会出现数据重新下载的情况"* type=*"int"*>12</property>

<property key=*"folder.record.file.regex"* desc=*"文件夹记录文件正则表达式 "* type=*"string"*>FolderAlreadyDownloadRecord-(\d{8})$</property>

<property key=*"connection.timeout.retry.times"* desc=*"连接超时重试次数"* type=*"int"*>3</property>

***(b)预处理文件输出配置（参照：config.xml）***

<property key=*"scanPeriod"* desc=*"扫描本地源文件的周期（单位：秒）"* type=*"long"*>60</property>

<property key=*"nativePath"* desc=*"src目录"* type=*"String"*>E:/export/home/sca/pretreatment/data/</property>

<property key=*"nativeBakDir"* desc=*"已经完成文件备份目录"* type=*"String"*>E:/export/home/sca/pretreatment/bak/</property>

<property key=*"http"* desc=*"http协议输出目录"* type=*"string"*>E:/export/home/sca/pretreatment/out/http-wap1x</property>

<property key=*"wap1x"* desc=*"wap1x协议输出目录"* type=*"string"*>E:/export/home/sca/pretreatment/out/http-wap1x</property>

<property key=*"conn"* desc=*"conn协议输出目录"* type=*"string"*>E:/export/home/sca/pretreatment/out/conn</property>

<property key=*"dns"* desc=*"dns协议输出目录"* type=*"string"*>E:/export/home/sca/pretreatment/out/dns</property>

<property key=*"mms"* desc=*"mms协议输出目录"* type=*"string"*>E:/export/home/sca/pretreatment/out/mms</property>

<property key=*"pdp"* desc=*"pdp协议输出目录"* type=*"string"*>E:/export/home/sca/pretreatment/out/pdp</property>

<property key=*"allProtocols"* desc=*"所有协议类型"* type=*"string"*>http,wap1x,conn,dns,pdp,mms</property>

<property key=*"tempFileExtension"* desc=*"临时文件扩展名"* type=*"String"*>.tmp</property>

<property key=*"finishFileExtension"* desc=*"完成文件扩展名"* type=*"String"*>.dat</property>

<property key=*"supplementKey"* desc=*"补充文件文件夹关键字"* type=*"String"*>supplement</property>

<property key=*"changeDirTime"* desc=*"修改目录tmp-->dat"* type=*"long"*>4800</property>

<property key=*"fileNameRegex"* desc=*"日志文件名称匹配正则"* type=*"String"*>\d{2}-\d{2}-\w+-(\w+)-\d{2}\_(\d{10})(\d{1})(\d{1})(\d{2})\_\d{2}$</property>

<property key=*"destDirNameRegex"* desc=*"目标目录名称正则"* type=*"string"*>\d{10}-\w+[-\w]\*</property>

<property key=*"isChangeTmpDirName"* desc=*"是否可以修改tmp文件夹名称，只要一台主机修改即可"* type=*"boolean"*>true</property>

<property key=*"tmpTransferDir"* desc=*"临时目录"* type=*"string"*>E:/export/home/sca/pretreatment/transfer-tmp</property>

<property key=*"overtime.file.config.path"* type=*"String"*>E:/eclipse/workspace/config/overtime.xml</property>

<property key=*"tascluster"* desc=*"大数据平台hadoop相关配置文件"* type=*"string"*>core-site.xml,hdfs-site.xml</property>

<property key=*"scacluster"* desc=*"SCA hadoop相关配置文件"* type=*"string"*>core-site.xml,hdfs-site.xml</property>

## 源数据说明

1. 在数据源中的数据为移动公司服务器产生的日志文件，目前共有两种，描述如下：

* 数据源文件：

01-01-gn-wap1x-02\_20130605155000\_00

01-02-gn-http-00\_20130605155000\_00

01-03-gn-dns-02\_20130605155500\_00

01-05-gn-mms-02\_20130605155500\_00

01-06-gn-conn-01\_20130605155000\_00

01-08-gn-pdp-00\_20130605155500\_00

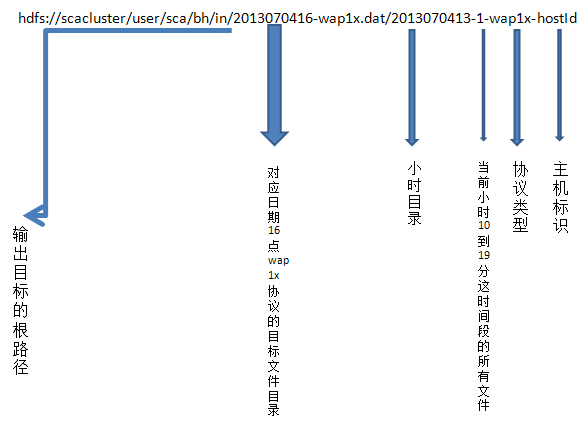
此类文件的文件名中包含两项关键要素：

1. 协议名——共有如下类型：wap1x、dns、http、mms、conn、pdp
2. 时间——如20130605155500，分解即为2013年0605日15时55分00秒

对此类文件的处理：程序将提取出协议名和时间（精确到小时）组成分组名，

## 输出数据说明

如下图所示



例子

01-08-gn-pdp-00\_20130605155500\_00

01-08-gn-pdp-00\_20130605155504\_00

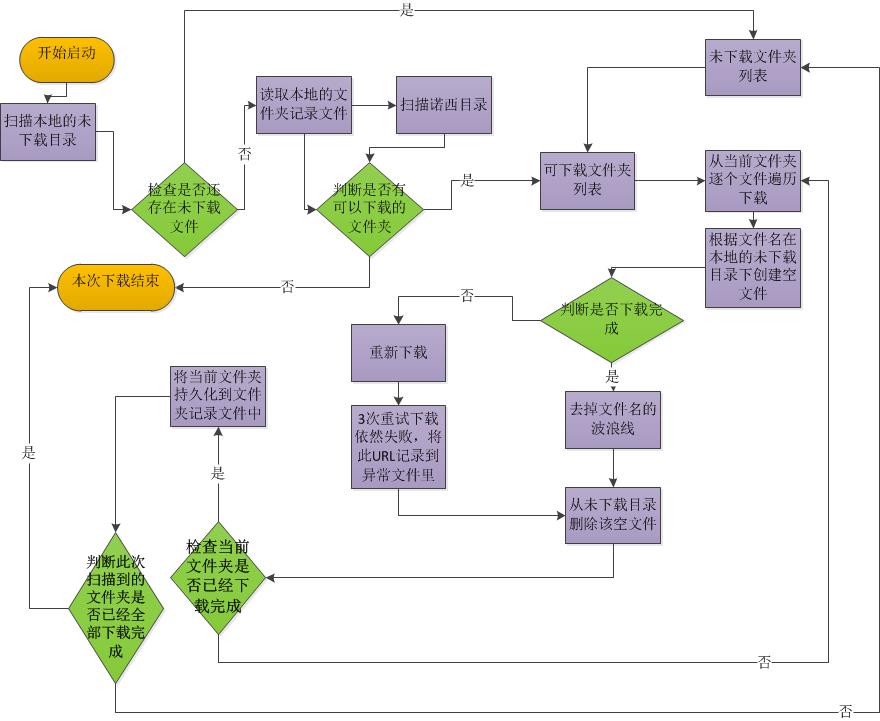
将被合并到2013060515-pdp组下的2013060515-5-pdp-hostId文件

01-08-gn-pdp-00\_20130605153500\_00

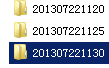
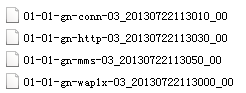
01-08-gn-pdp-00\_20130605153504\_00

将被合并到2013060515-pdp组下的2013060515-3-pdp-hostId文件

预处理模块文件采集流程图



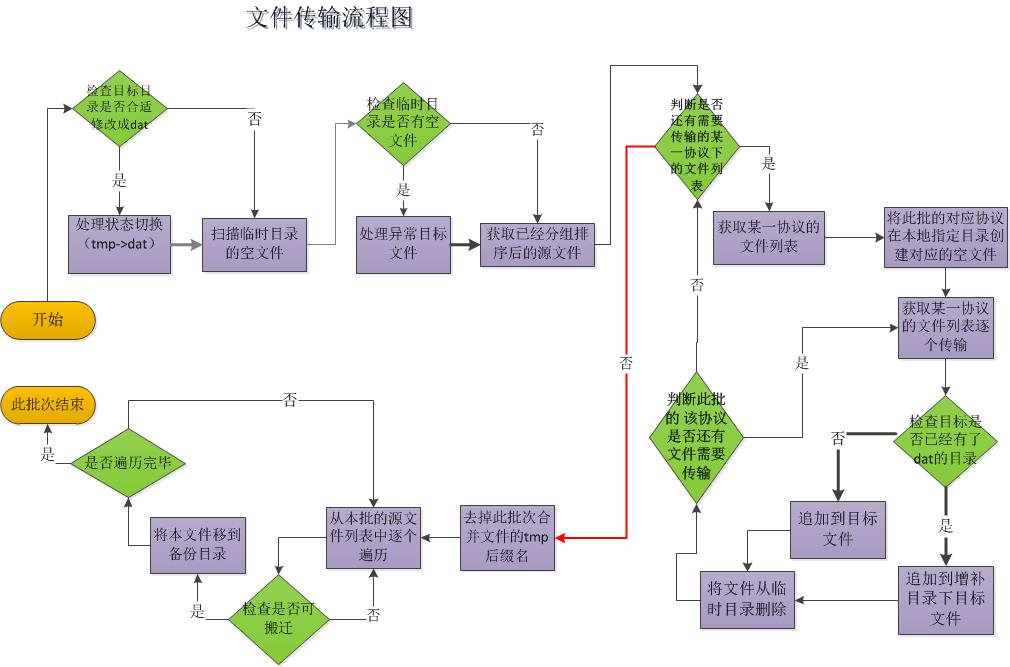
数据采集模块程序将每隔5分钟（可配置” *scan.time.interval.value*”）扫描指定的远程数据源根路径，获取诺西15（可配置）分钟以前所有的文件夹，然后与本地记录已经下载过的文件夹相比较获得已经下载过的文件夹的补集，也就是待下载文件夹列表，接着遍历诺西的文件夹列表将其里面的文件并行下载到采集机的指定目录下。

示例： 

比如对“201307221130”这个文件夹下的文件列表如上图，文件采集的处理流程如下

1. 扫描本地的未下载目录，将其没有下载的文件重新下载下来
2. 扫描FTP源文件存放的根目录，获取该目录下所有的文件夹列表
3. 读取本地的文件夹记录文件获取已经下载过的文件夹列表
4. 将第二步的文件夹列表和第三步的文件夹列表进行比较得出待下载的文件夹列表
5. 对未下载的文件夹列表进行过滤只取15分钟以前生成的文件夹
6. 通过a-d步骤之后得到的文件夹列表，逐个文件夹进行遍历下载文件夹里面的文件
7. 在下载之前把当前待下载的文件夹里的所有文件在本地的一个临时目录下创建对应的一个空文件，当把这个文件下载完成之后，将其删除掉。
8. 检查当前文件夹是否已经下载完成，如果下载完成将其文件夹名称记录到文件夹记录文件

## 预处理模块文件输出流程图



1. 根据设定的时间修改目标文件的状态，将.tmp为后缀的文件改成以.dat为后缀
2. 扫描临时目录处理异常目标文件
3. 获取已经分组排序后的文件列表
4. 按协议遍历所有的文件列表
5. 在本地指定目录下创建对应的空文件
6. 检查目标目录下是否已经存在对应的dat目录，如果存在追加到增补目录下的目标文件，如果不存在追加到目标文件
7. 传输完成之后将其文件从临时目录下删除掉
8. 循环a—>g步骤，直到遍历完为止
9. 将此扫描的文件传输完成之后把目标文件的.tmp后缀去掉
10. 接着将本地的源文件移到备份目录下

注：这里的tmp有两种:

1、目标文件上一级目标由tmp转变成dat，为tmp表示该目录下的合并文件还需要继续往该目录下写数据，dat表示传输合并文件已完成，并不再往该目录下写数据

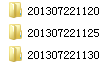
2、目标文件有一个状态为tmp

a>相对某一批次来说，后缀为tmp则可以继续往该文件追加数据

b>当下一次扫描发现目标目录下还有tmp为后缀的目标文件，则说明该文件是合并过程中出现了异常，所以该文件需要删除掉

## 其他细节处理逻辑

1. 输出目录的某个分类文件夹如2013071915-mms.dat 目录下有小于等于24个文件，在该目录下的文件的规则为“yyyyMMddHH-(批次0-5)-(协议)-主机的ID”，如：2013072215-5-mms-1374820609777 ，有四台采集机，每小时每一协议会产生小于等于6个目标文件
2. 在每一个传输周期开始都要去判断目标目录下检查是否存在.tmp的目标文件，如果存在将其删除掉（此步处理的来由：目标文件在生成的过程中可能出现失败，如果出现失败，那目标目录下会存在对应以tmp为后缀的文件，从而在目标路径下留下一个残留的合并文件。上述处理可以解决这个问题）
3. 当扫描发现本次输出的目录已经是.dat，那此时会在与.dat目录创建一个增补目录，将延迟到达的文件合并写到增补目录下，其命名规则与正常流程一致
4. 在采集过程中，发现当前文件夹下有波浪线为结尾的文件，此时也会把该文件下载下来，但此文件不会放到本地的源目录下，而是放到一个异常文件的目录下以备核对异常文件时用
5. 文件采集流程中，FTP上的目录与目录下的文件可以不在同一时间段
6. 诺西每5分钟生成一个目录文件，其目录结构为：



## 历史数据备份管理

a)以1小时为周期，定时扫描备份目录，当发现存在超过配置文件里配置的时间值的目录将其删掉，定时清理备份目录，确保备份目录下只存放最近的备份文件。

<bean>

<protocol>http</protocol><!-- 协议的类型 -->

<switch>on</switch><!-- 扫描开关 -->

<basePath>E:/export/home/sca/pretreatment/bak</basePath><!—备份文件的根路径 -->

<scanCycle>1</scanCycle><!-- 扫描周期以小时为单位 -->

<scanTargetRegex>\d{10}-\w+[-\w]\*</scanTargetRegex><!-- 扫描目录的正则匹配表达式 -->

<timeRegex>\.\*?\_(\d{4})(\d{2})(\d{2})(\d{2})</timeRegex><!-- 扫描目录的时间元素抽取正则表达式 -->

<deadline>48</deadline><!-- 备份文件的存储期限时间数值 -->

<timeUnit>hour</timeUnit><!-- 备份文件的存储期限时间单位可取 year month day hour minute second -->

</bean>

## 根据测试进度待补充