**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

二零一三年六月

**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 | 2013.6.11 | 1.0 | 创建文档 | 段海涛 |  |  |
| 2 | 2013.6.25 | 2.0 | 功能描述 | 段海涛 |  |  |
| 3 | 2013.6.26 | 2.01 | 添加清扫临时文件功能说明  修改诺西经分两类数据的识别方式 | 段海涛 |  |  |
| 4 | 2013.6.29 | 2.02 | 传输文件流程增加检查残余文件逻辑  修改采集失败后的重试方式  修改文件生成时间与采集时间的时间差配置，两种数据统一用小时为单位 | 段海涛 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

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 5](#_Toc360001227)

[3.1 配置文件 5](#_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 8](#_Toc360001231)

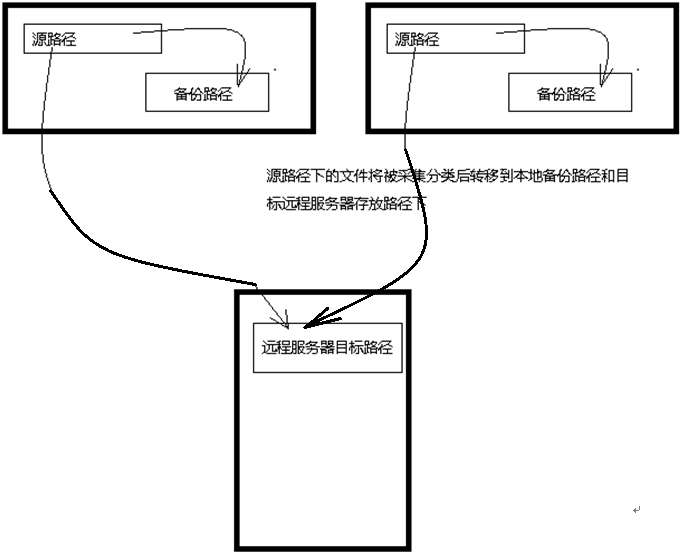
[3.5 根据测试进度待补充 8](#_Toc360001232)

**数据采集模块项目说明**

# 功能概述

本模块的主要功能是循环采集日志文件服务器上指定目录下的日志文件，并进行分类合并后转移到目标存储位置下的分类文件夹中（包括本地的备份目录和远程hadoop集群等）。

# 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

参数配置在config.xml文件中

配置示例如下：

<!-- 数据采集模块开关 -->

<property key=*"data.collect.switch"* type=*"string"*>on</property>

<!-- 数据采集扫描周期，单位为分钟 -->

<property key=*"data.collect.sycle"* type=*"string"*>1</property>

<!--每一个分组收满时的文件数 -->

<property key=*"max.file.count"* type=*"string"*>2</property>

<!--分组的更改状态的时限 -->

<property key=*"max.file.time"* type=*"string"*>1</property>

<!--划分分组时需要的步骤数 -->

<property key=*"level.group"* type=*"string"*>2</property>

<!--往分组文件夹传输文件时是否需要将本次传输组的文件合并 -->

<property key=*"data.collect.filecombine"* type=*"string"*>true</property>

<!--文件名分类字串截取正则表达式，根据实际抽取需要来配置 -->

<!-- sca -->

<property key="whole.regex" type="string">\d{2}-\d{2}-\w+-\w+-\d{2}\_\d{14}\_\d{2}</property>

<property key="first.level.regex" type="string">\_(\d{10})</property>

<property key="second.level.regex" type="string">\d{2}-\d{2}-\w+-(\w+)-</property>

<!-- 数据采集源数据路径 -->

<property key=*"data.collect.src"* type=*"string"*>/home/hadoop/logsrc</property>

<!-- 数据采集历史文件夹扫描路径，用来更改该路径下的过期tmp文件夹状态 -->

<property key=*"data.collect.scandir1"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/in/</property>

<property key=*"data.collect.scandir2"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/other/</property>

<!-- 经分日志采集后本地备份路径 -->

<property key=*"JFen.bak"* type=*"string"*>/home/hadoop/jfbak/</property>

<!-- wap1x协议日志采集后本地备份路径 -->

<property key=*"wap1x.bak"* type=*"string"*>/home/hadoop/logbak/</property>

<!-- http协议日志采集后本地备份路径 -->

<property key=*"http.bak"* type=*"string"*>/home/hadoop/logbak/</property>

<!-- dns协议日志采集后本地备份路径 -->

<property key=*"dns.bak"* type=*"string"*>/home/hadoop/logbak/</property>

<!-- mms协议日志采集后本地备份路径 -->

<property key=*"mms.bak"* type=*"string"*>/home/hadoop/logbak/</property>

<!-- conn协议日志采集后本地备份路径 -->

<property key=*"conn.bak"* type=*"string"*>/home/hadoop/logbak/</property>

<!-- pdp协议日志采集后本地备份路径 -->

<property key=*"pdp.bak"* type=*"string"*>/home/hadoop/logbak/</property>

<!-- 不符合分类规则的异常文件的本地转移备份路径 -->

<property key=*"imnormal.bak"* type=*"string"*>/home/hadoop/otherbak/</property>

<!-- 经分日志采集后转移到的远程目的 -->

<property key=*"JFen.dest"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/jfen/</property>

<!-- wap1x协议日志采集后转移到的远程目的 -->

<property key=*"wap1x.dest"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/in/</property>

<!-- http协议日志采集后转移到的远程目的 -->

<property key=*"http.dest"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/in/</property>

<!-- dns协议日志采集后转移到的远程目的 -->

<property key=*"dns.dest"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/other/</property>

<!-- mms协议日志采集后转移到的远程目的 -->

<property key=*"mms.dest"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/other/</property>

<!-- conn协议日志采集后转移到的远程目的 -->

<property key=*"conn.dest"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/other/</property>

<!-- pdp协议日志采集后转移到的远程目的 -->

<property key=*"pdp.dest"* type=*"string"*>hdfs://192.168.193.130:9000/sca/bh/other/</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：

* 第一类数据源文件：

MASA\_USER\_20130616\_001.txt

TB\_DIC\_AC\_ACCT\_ITEM\_20130616\_001.txt

TB\_DIC\_AREA\_CELL\_NEW\_20130616\_001.txt

TB\_DIC\_AREA\_CODE\_20130616\_001.txt

TB\_DIC\_BRAND\_20130616\_001.txt

对此类文件的处理：程序将提取出时间（精确到日期）生成分组名，如:

文件： TB\_DIC\_AREA\_CODE\_20130616\_001.txt

TB\_DIC\_AREA\_CODE\_20130616\_001.txt

MASA\_USER\_20130616\_001.txt

都将被分组到20130616组

* 第二类数据源文件：

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，如:

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

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

都将被分到pdp-2013060515组

注：以上两类文件不会在一个实际采集系统中同时出现；

程序也不能同时处理这两类文件；

但可通过相关正则表达式的配置，程序可自动识别处理的是哪种类型数据；

*a)经分数据的正则配置：*

*<property key="whole.regex" type="string">[\w\_]\*?\_(\d{8})\_[\w\_]\*?.\*?</property>*

*<property key="order.regex" type="string">1</property>*

*<property key="key.regex" type="string">(\d{8})</property>*

*<property key="tmp.regex" type="string">(\d{8})\.tmp</property>*

*b)诺西数据的正则配置：*

<property key="whole.regex" type="string">\d{2}-\d{2}-\w+-(\w+)-\d{2}\_(\d{10})\d{4}\_\d{2}</property>

<property key="order.regex" type="string">12</property>

<property key="key.regex" type="string">(\w+)-\d{10}</property>

<property key="tmp.regex" type="string">.\*?(\d{10})\.tmp</property>

## 数据处理主体流程：

模块程序将每隔20分钟（可配置”*data.collect.sycle*”）扫描配置文件中指定的数据源路径下如上图所示的文件，并按照文件名中的协议和时间进行分类；

示例：比如对“01-01-gn-wap1x-02\_20130605155500\_00”这个文件的处理流程

1. 扫描检测配置文件中预设的数据源目录下的所有文件，进行下述遍历分组处理：
2. 根据配置好的正则表达式截取文件名的协议和时间：wap1x和2013060515
3. 将上述截取到的两个字串合并成 wap1x-2013060515
4. 在内存中将本次扫描获取的文件按上述规则分成若干个文件组；

如组wap1x-2013060515

wap1x-2013060516

http-2013060515

……

1. 在采集的目标根路径按 ”组名.tmp” 格式建立分组文件夹

将属于某分组（如wap1x-2013060515）的所有文件：

如01-01-gn-wap1x-02\_20130605155500\_00

01-01-gn-wap1x-02\_20130605155500\_02

……

合并存入远程目标目录下的wap1x-2013060515.tmp文件夹内；

生成一个临时统计文件用于计数；

然后将源数据目录的这组文件全部原样转移到本地的备份目录下（可配置）

1. 待一组文件处理完毕后，程序会检查分组文件夹wap1x-2013060515.tmp内接收过的文件数量，如果达到配置文件中配置的上限数量，程序会将wap1x-2013060515.tmp更名为wap1x-2013060515.dat

而后，程序继续循环处理下一组文件；

1. 在一次扫描处理周期结束前，程序还会检查配置文件中配置的需要做“过期”监控的路径（可配置），找出带.tmp后缀且已经过期（过期时限可配置）的分类文件夹，根据配置对该文件夹内的number-random.count文件进行清理，然后将文件夹后缀名更名为.dat

## 其他细节处理逻辑

1. 当某个分类文件夹如wap1x-2013060515.tmp下收集到的文件已满60（可配置）个，则该文件夹会被程序自动改名为wap1x-2013060515.dat ；
2. 在传输一组文件之前，会先判断目标路径下是否已经存在该文件组，如果存在，则将目标文件夹中的同名合并文件删除，然后再重建并写入数据；

（此步处理的来由：一组文件在传输过程中可能出现失败，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）

1. 当某个分类文件夹在一定时限后（可配置）还未能收集满60个文件，则该文件夹也会被改名，即在原来的名字后加上后缀 “.dat” ；
2. 当扫描到的某文件所归属的分组所对应的文件夹在此前已被打上“.dat”后缀，该文不会输出到所配置的远程目标和本地备份文件夹内；而是直接转移到一个other（可配置）文件夹中；
3. 当扫描到的文件的文件名不符合正常格式，将会被作为异常文件处理，即直接转移到异常文件夹中（可配置）；
4. 扫描线程的扫描频率为20分钟一次（可配置），在扫描过程中，如果发生源数据读取异常或目标路径写入异常，程序将重试，重试时间间隔为2秒；

如果重试成功，则又恢复20分钟循环周期；

1. 本程序能适应多台采集服务器往同一个目标路径下输出分类结果的需求；
2. 本程序有一个总的开关，*"data.collect.switch"，*

*值为OFF时，本模块将不会运作*

*值为ON时，本模块正常工作*

## 清扫已完成文件分组文件夹内的临时校验文件

a)功能来由：为了统计分组文件夹所接收的文件数量采用的一种处理方法

b)功能流程详述：

每一台采集服务器往目标路径输送文件的时候，会将一次扫描中扫描到的该组文件合并成一个文件

为了便于统计该分组文件夹中共计接收过的文件数，每台采集服务器在一次扫描传输中会在该分组文件夹下写一个用于统计的临时文件。

该临时文件的命名规则: number-random.cout

number：本次传输的文件个数

random：随机数

当统计到该分组已接收的文件数量达到配置中指定的文件数量时，程序将清除之前的临时文件（可配置）；

c)功能预期结果：

在每一个tmp状态分组文件夹中，会存在一些number-random.cout 临时文件

在每一个dat状态分组文件夹中，这些.count临时文件将不复存在

d)该功能有一个配置开关:

<property key="data.collect.ifclean" type="string">true</property>

配为true时，清扫功能开启

配位false时，清扫功能关闭

## 根据测试进度待补充