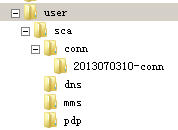
**预处理模块输出接口说明**

一、输出目录结构：

|  |
| --- |
| hdfs://tascluster/user/sca/conn/20130704-conn/2013070413-conn-combinedBy-hostId-log |
| hdfs://tascluster/user/sca/dns/20130704-pdp/2013070413-pdp-combinedBy-hostId-log |
| hdfs://tascluster/user/sca/pdp/20130704-dns/2013070413-dns-combinedBy-hostId-log |
| hdfs://tascluster/user/sca/pdp/20130704-dns/2013070413-dns-combinedBy-hostId-log |
| hdfs://scacluster/user/sca/bh/in/20130704-http/2013070413-http-combinedBy-hostId-log |
| hdfs://scacluster/user/sca/bh/in/20130704-wap1x/2013070413-wap1x-combinedBy-hostId-log |



二、输出文件规则

* 每一台采集机将同一小时时段内的同一类型日志文件合成一个文件输出到目标目录
* 输出过程中，输出文件夹会带上.tmp后缀，如：

hdfs://tascluster/user/sca/conn/20130704-conn.tmp

* 输出完成后，输出文件夹后缀会变更为.dat，如：

hdfs://tascluster/user/sca/conn/20130704-conn.dat

* 输出文件的周期：

每10分钟一次往合并文件中追加内容

每小时完成一个合并文件