中国移动采购和招标网数据爬取程序——数据存储说明以及与网站内容对应信息

1. **基本信息：**

|  |  |
| --- | --- |
| 存储介质 | hbase集群(181) |
| zookeeper | n1.cluster,n2.cluster,n3.cluster |
| namespace | ns1 |
| table结构 | info2(采购公告),info7(结果公示) |
| family | ‘base’,’extend’ |
| fields | [见下方,数据类型都是字符串](#_fields信息) |
| rowkey | id.hashcode+””+date.hashcode |

1. **详细说明：**

1). xwtech用户无法创建namespace,没有权限，hbase表需要预先创建

create ‘ns1:info2’,{NAME=>’base’}

create ‘ns1:info7’,{NAME=>’base’},{NAME=>’extend’}

1. . 如果需要更换hbase集群，修改/home/xwtech/chenxiang/configuration.properties中

hbaseZookeeperQuorum=n1.cluster,n2.cluster,n3.cluster

1. . table表名由”info”+noticetype构成（noticetype是中国移动采购和招标网发布的信息内容的标识,对应关系如下）,现在程序只执行info2(采购公告),info7(结果公示)。

采购公告 2

资格预审公告 3

结果公示 7

单一来源采购信息公示 1

供应商信息收集公告 8

1. . 有两个family，分别为base和extend,base存储一级页面的内容，extend存储二级页面的内容（只在结果公示中使用）

base需要在每张表中都创建好。

extend只需要在info7（结果公告里创建即可）

## fields信息

|  |  |  |
| --- | --- | --- |
| base | name | 采购需求单位 |
| type | 公告类型 |
| content | 标题 |
| date | 时间 |
| id | 二级页面编号 |
| extend | title | 结果公示标题 |
| content | 公示内容 |



