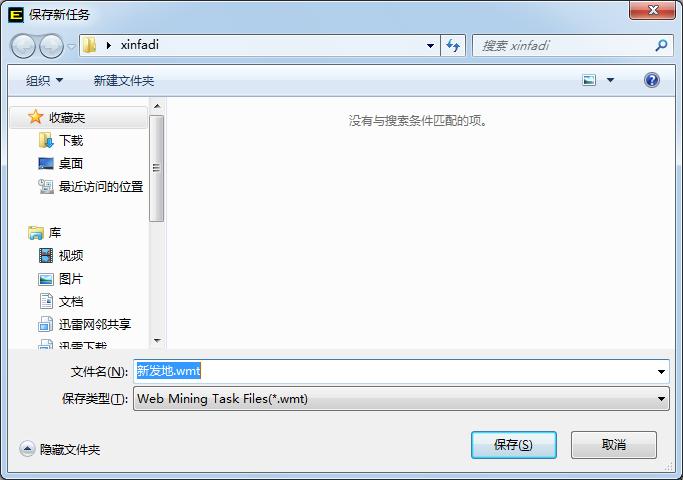
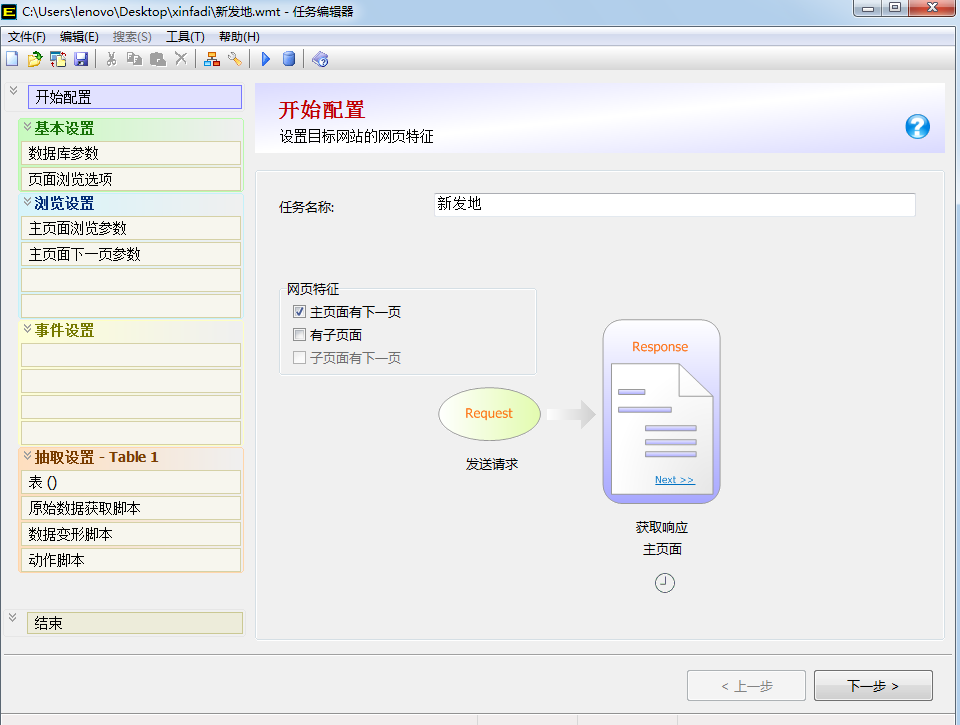
时间：2015年10月27日

地点：二教303

讨论内容：演示如何抓取新发地品种属性信息

正文：

1. 打开软件，点击“配置任务”，将任务保存到新发地文件夹中。
2. 新发地（<http://www.xinfadi.com.cn/marketanalysis/0/list/1.shtml>）上有下一页，因此网页特征点击主页面有下一页。

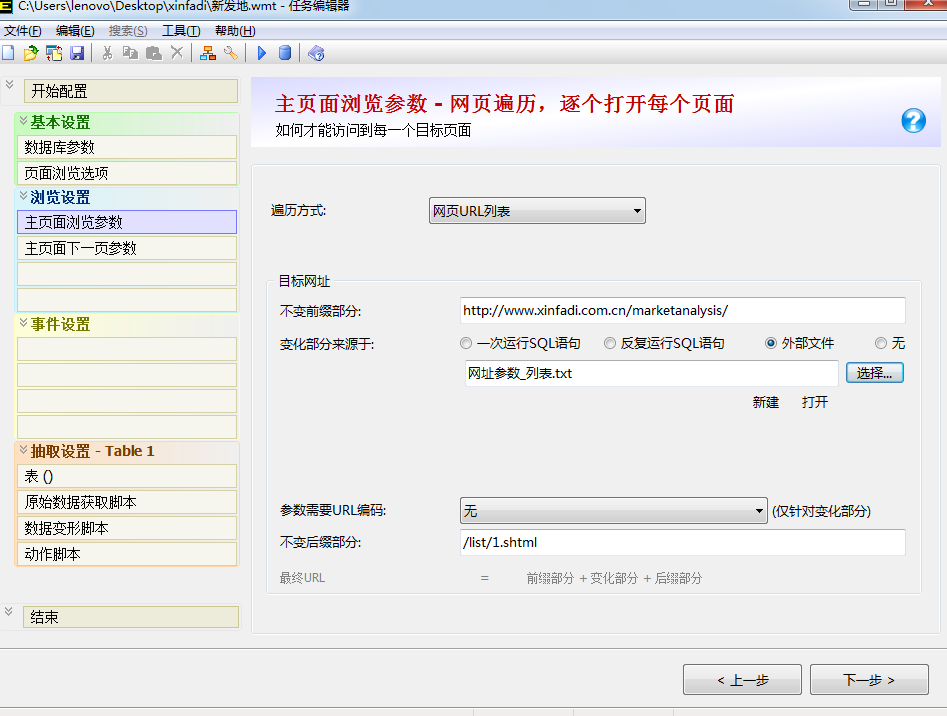
（3） 数据库参数设置默认即可，页面浏览设置默认即可。

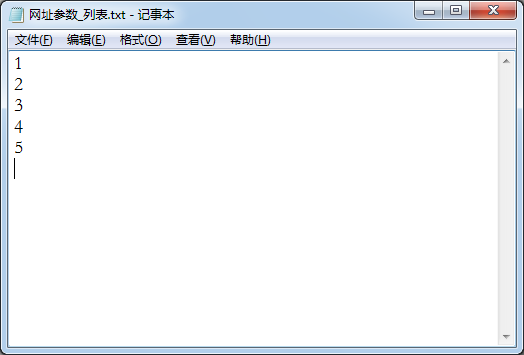
（4） 主页面浏览参数设置：

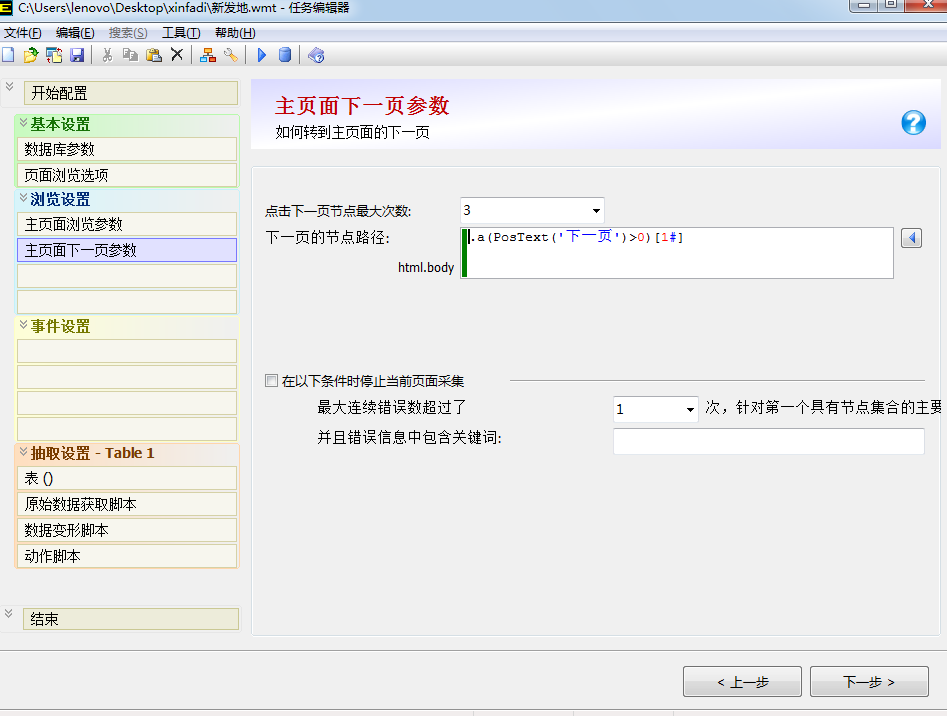
观察到新发地页面点击蔬菜一栏时对应的网址中数字是1（红色箭头处），而点击水果时对应网址中的数字是2，依此类推。



因此如果要一次性抓取蔬菜，水果，肉禽蛋，水产，粮油的数据，就必须设置浏览器网址，将<http://www.xinfadi.com.cn/marketanalysis/>放在不变前缀部分，变化部分是数字1,2,3,4,5，将其放置在外部文件，点击打开外部文件，将里面内容修改为如下格式，将/list/1.shtml放置在不变后缀部分。这三部分组合起来是一个完整的网页地址。



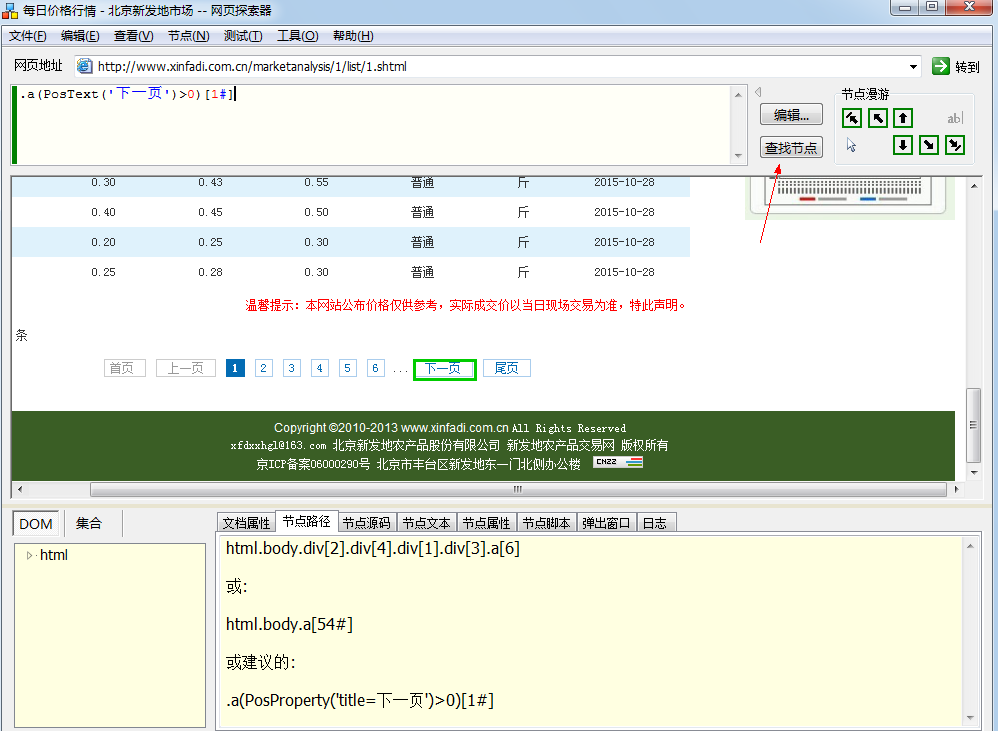
（5）主页面下一页参数中我们选择抓取前4页，选择3，软件实际抓取要多一页。下一页节点路径点击右侧三角，就会出现一个节点路径。



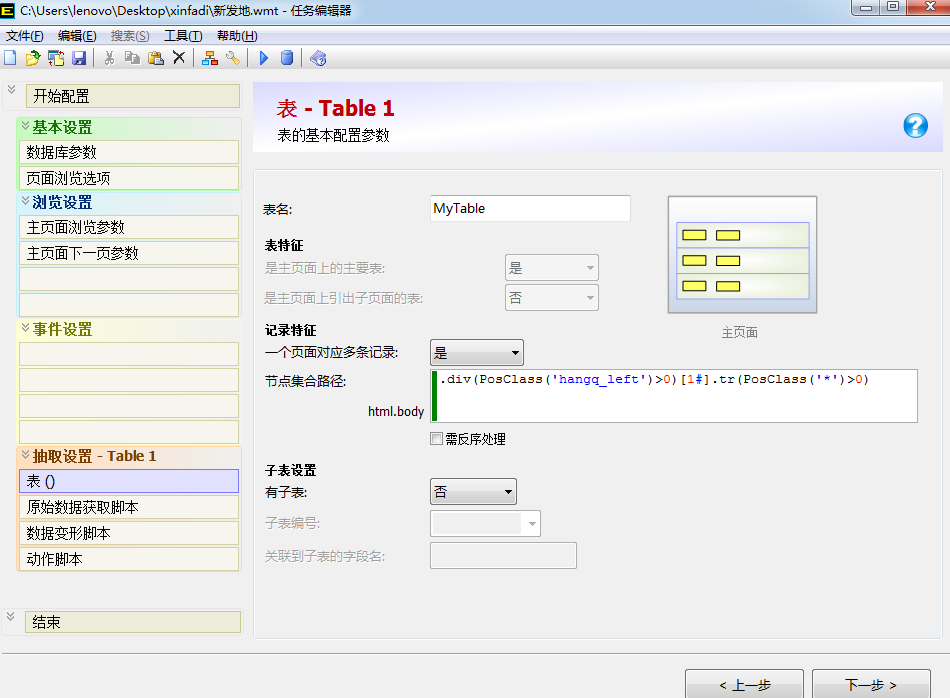
点击分析网页，出现网页探索器，



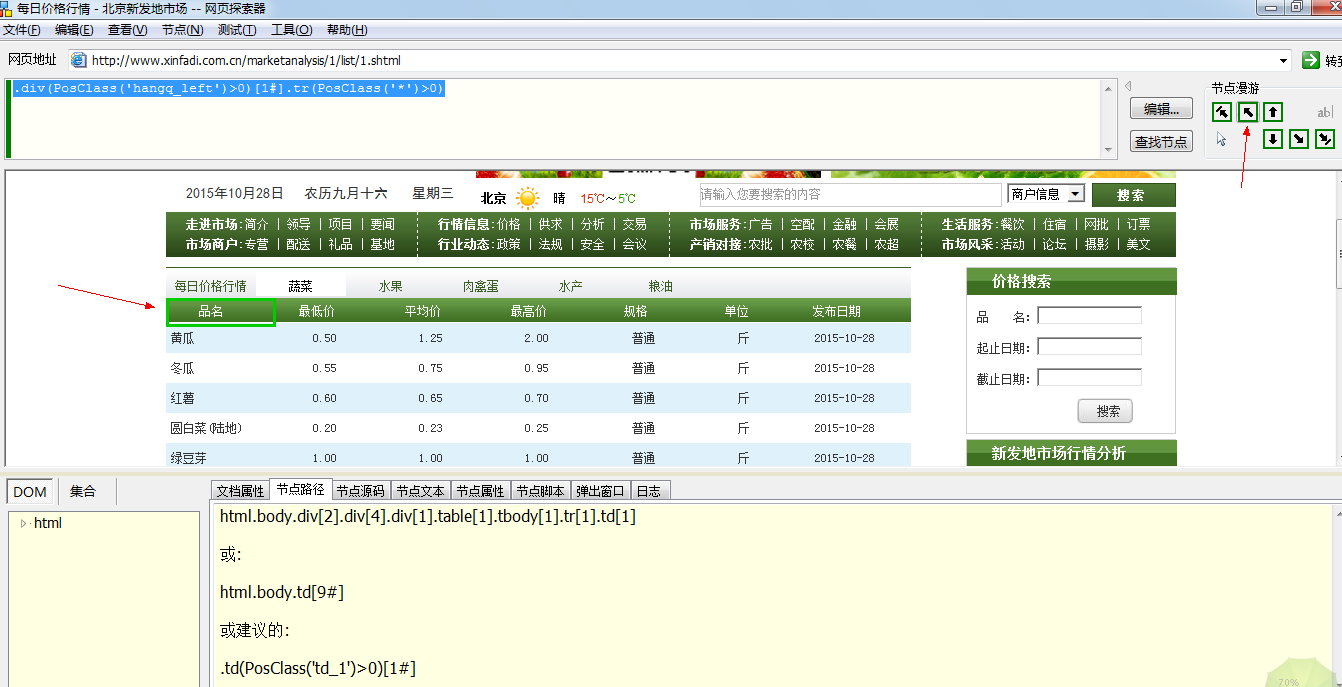
在网页探索器中的网页地址处输入待抓取网页网址，点击转到，就会出现待抓取网页。之后将主页面下一页参数中的下一页节点路径复制到网页地址下的方框中，如下图，之后点击查找节点，查看绿色光标是否定位在网页中的下一页，如果定位正确，说明节点路径正确可用。本例中正确。

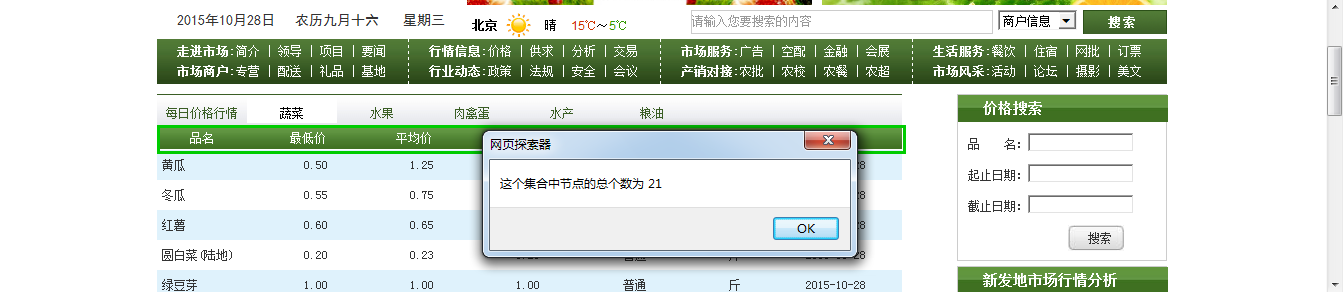


（6）之后进入到表的配置，本页有21行观测值，因此一个页面对应多条记录选择是，节点集合路径通过网页探索器来寻找。



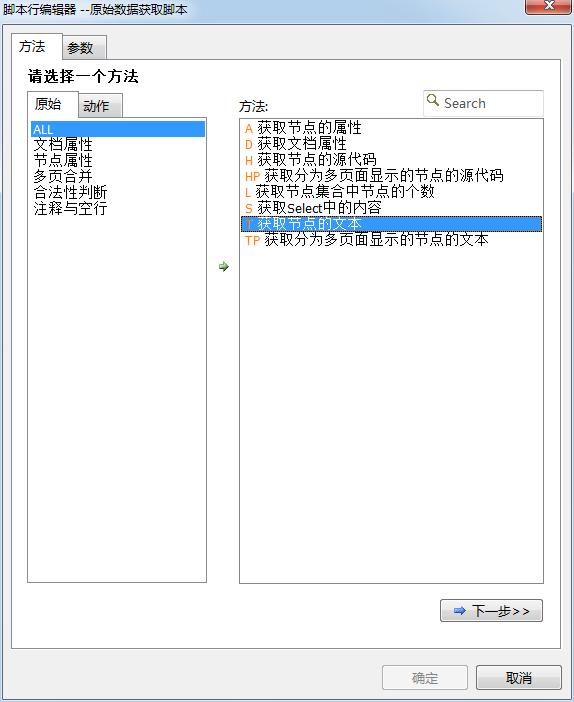
在网页探索器中，随便点击数据框中的一个位置，如品名，之后点击右上角的父节点，直到把所有数据都包含在内为止。查看下方的节点路径，.div(PosClass('hangq\_left')>0)[1#]，将路径复制到上方的框中，之后点击品名，点击父节点，将这一行所有数据都包含在内，查看此时的节点路径: .tr(PosClass('tr\_1')>0)[1#]，其中tr代表数据框的行，[1#]代表第一行，将单引号内的tr\_1变为通配符\*,表示所有，将节点路径添加到网页地址下方的框中，去掉[1#]，之后点击查找节点，出现节点数为21个，此页中共有21行数据，判断正确，表示路径可用。





（7）

在原始数据获取脚本中，添加变量，我们要获取数值，采用获取节点文本。

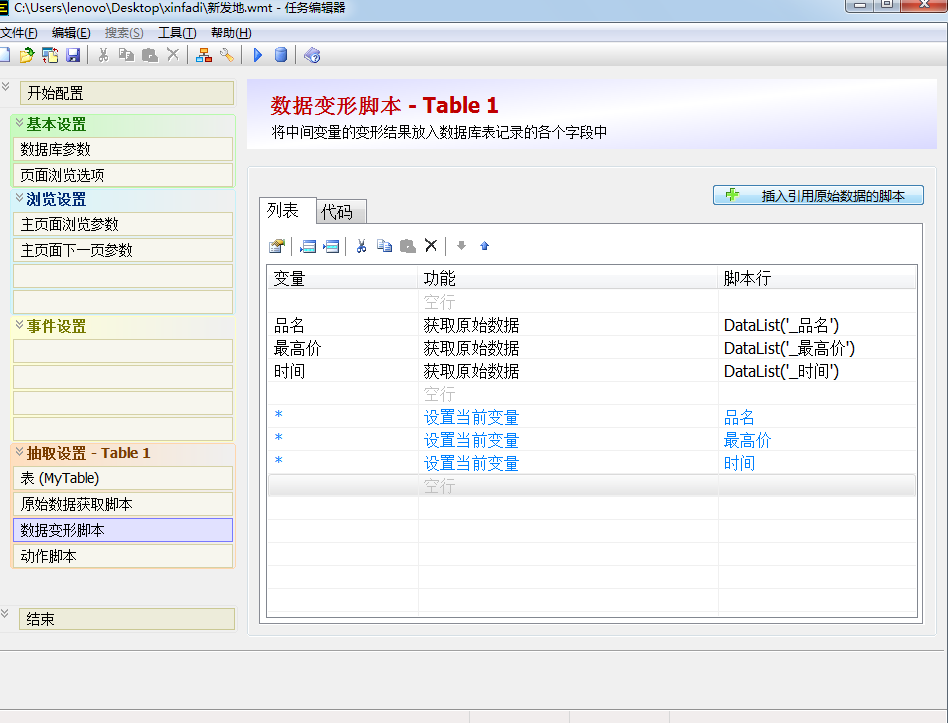


输入变量，格式为\_变量名，节点路径采用相同的方法，注意务必使每一行都可以采用同样的方式找到节点。下面的路径同样采用了通配符\*。



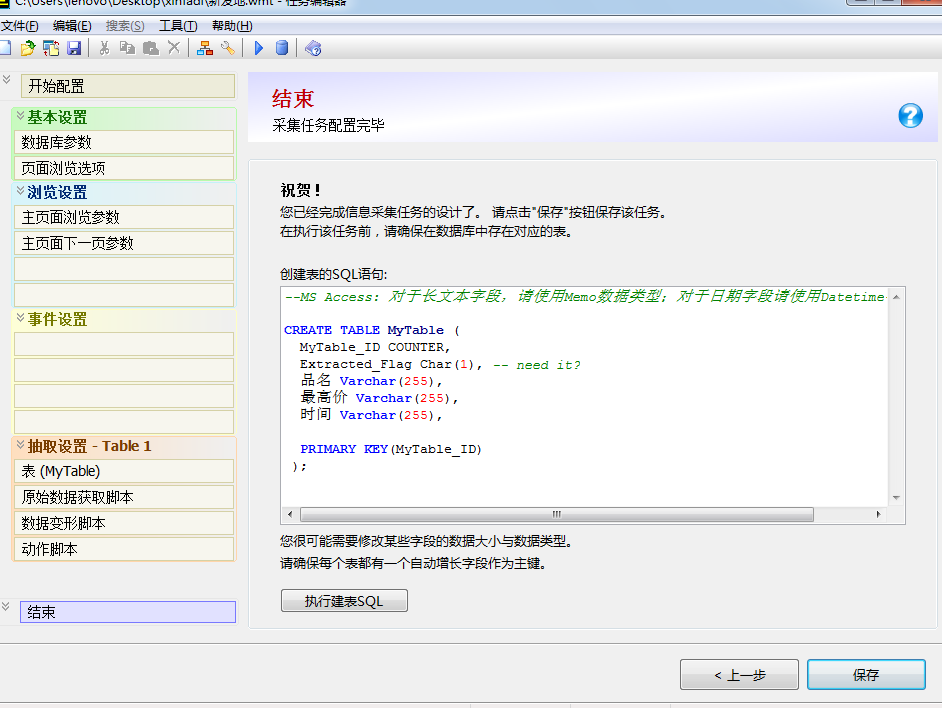
采用同样的方法获取到三个变量，品名，最高价，时间。

（8）点击数据变形脚本，点击右上角的插入引用原始数据的脚本，结果如下。

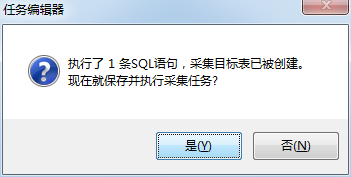


（9）动作脚本根据需要设置，本例中无须设置。

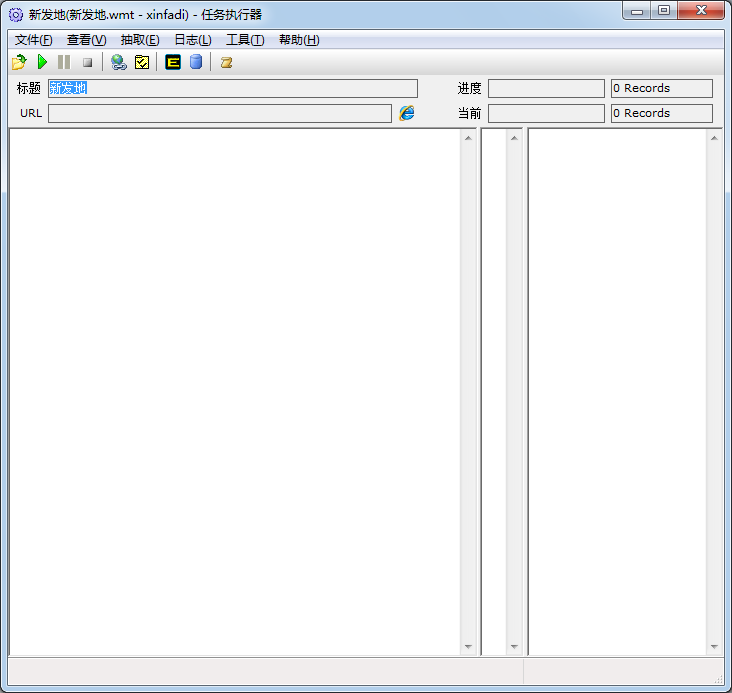
（10）点击结束，点击执行建表SQL



如果设置正确，会出现如下提示，点击是。



（11）出现如下界面，点击左上角的绿色前进标志，执行程序，

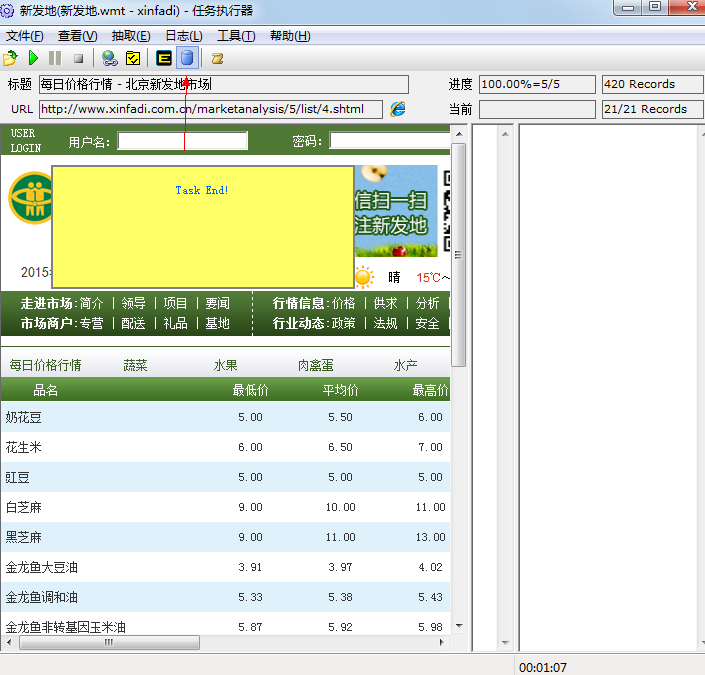


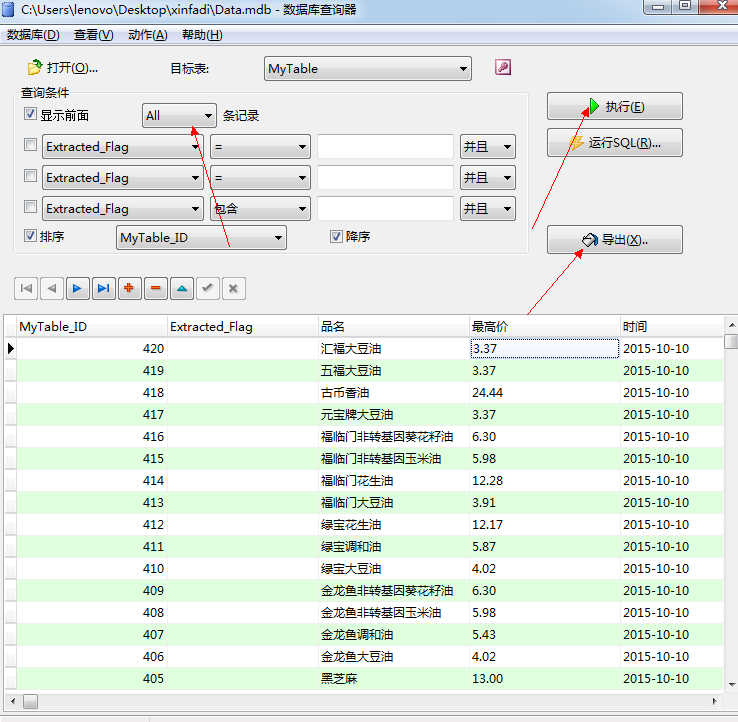
运行过程中会出现如下界面，右上角表示数据采集进程。



（12）任务完成后，点击箭头指向处，数据库查询。

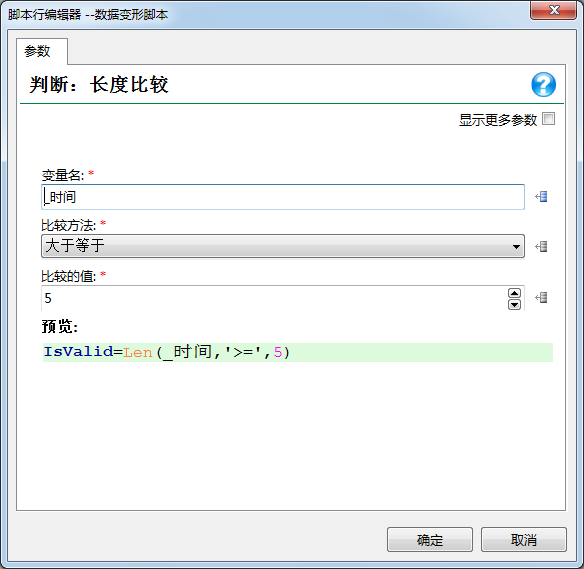
选择显示数量，点击执行，显示最终的数据结果，一页有85个数据，总共有4页，共420个数据。采集成功！



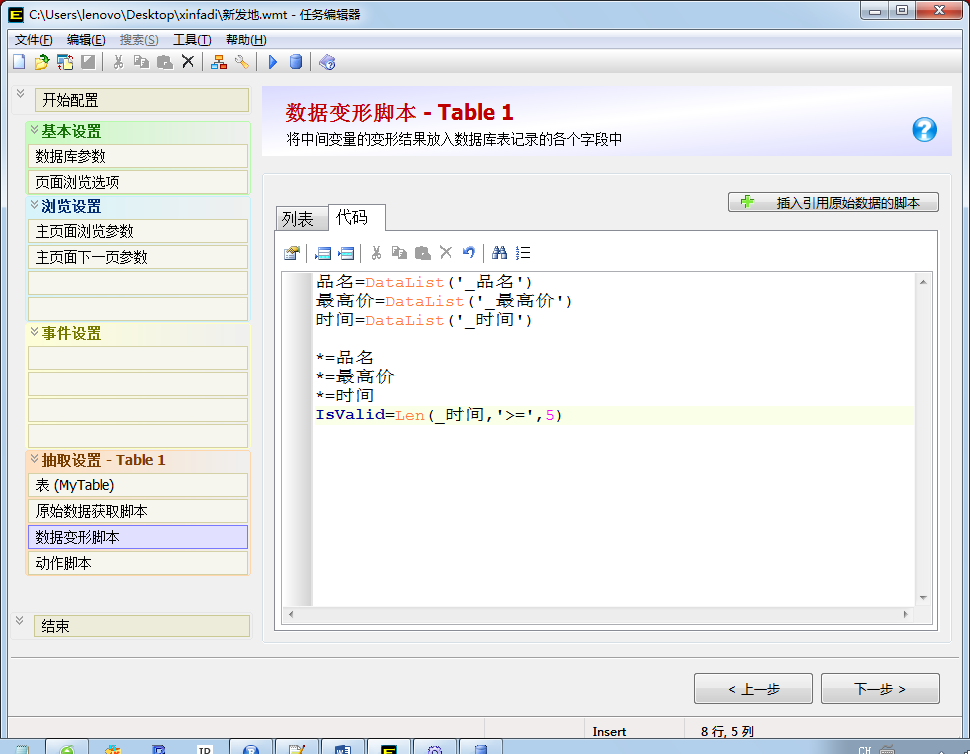


附注：

如果要滤掉第一行的标题，可以在数据变形脚本中采用合法性判断，比如判断时间长度小于5的扔掉。



将判断后的变量放在\*=时间后面，如下图。



之后的操作同之前一样，但是可以实现将第一行标题滤掉的效果。