

# CONBY C3 运营平台

## Map-Reduce 数据采集应用设计指南

### Version 1.0

高性能超大规模业务流程自动化引擎  
高性能大规模搜索引擎式数据挖掘引擎  
高性能大规模规则式商业智能引擎  
高性能大规模规则式社交网络智能引擎  
.....

技术代码: C3V3

品牌名称: C3

商标设计:



## 版权声明

本文档是上海创百信息技术有限公司(*Conby Information Technology Co., Ltd.*)针对 CONBY C3 运营平台 (CONBY C3 Operating Platform) 发布的产品白皮书, 未经上海创百信息技术有限公司同意, 任何组织或个人不得转载、修改、引用, 或做任何未经许可的用途。

上海创百信息技术有限公司保留对所有侵权行为提出法律诉讼的权利。

本文中的“CONBY”, “CONBY C3 运营平台”, “CONBY C3 Operating Platform” 为上海创百信息技术有限公司的专有名称, 其余专门术语和缩写分属相应国际组织和公司。

上海创百信息技术有限公司 平台运营事业部

电话: 020-38293726/606

传真: 020-38296095

邮件: [market@conby.com](mailto:market@conby.com)

网址: [www.conby.com](http://www.conby.com)

地址: 中国广州市珠江新城 CBD 星汇国际大厦东塔 1102 室

## 目录

一、入门实例.....	4
1、用户注册、登录.....	4
2、数据源搜索规则定义.....	6
3、数据源Task定义.....	8
4、Map-Reduce计算Task定义.....	11
5、Map计算Task定义.....	12
6、Reduce计算Task定义.....	13
7、Map-reduce计算过程输出.....	14
8、离散事件系统流程图.....	16
二、Map-Reduce计算架构图.....	17
三、模拟浏览器自动登录并发微博.....	18
1、微博消息源定义.....	18
2、人工微博消息事件定义定义.....	19
3、微博登录及发送规则定义.....	20
4、微博登录及自动发微博Task.....	22
5、激活消息源触发登录及自动发微博动作.....	23
6、触发人工事件引发自动发微博动作.....	23
7、检查微博验证结果.....	24
8、离散事件系统流程图.....	24

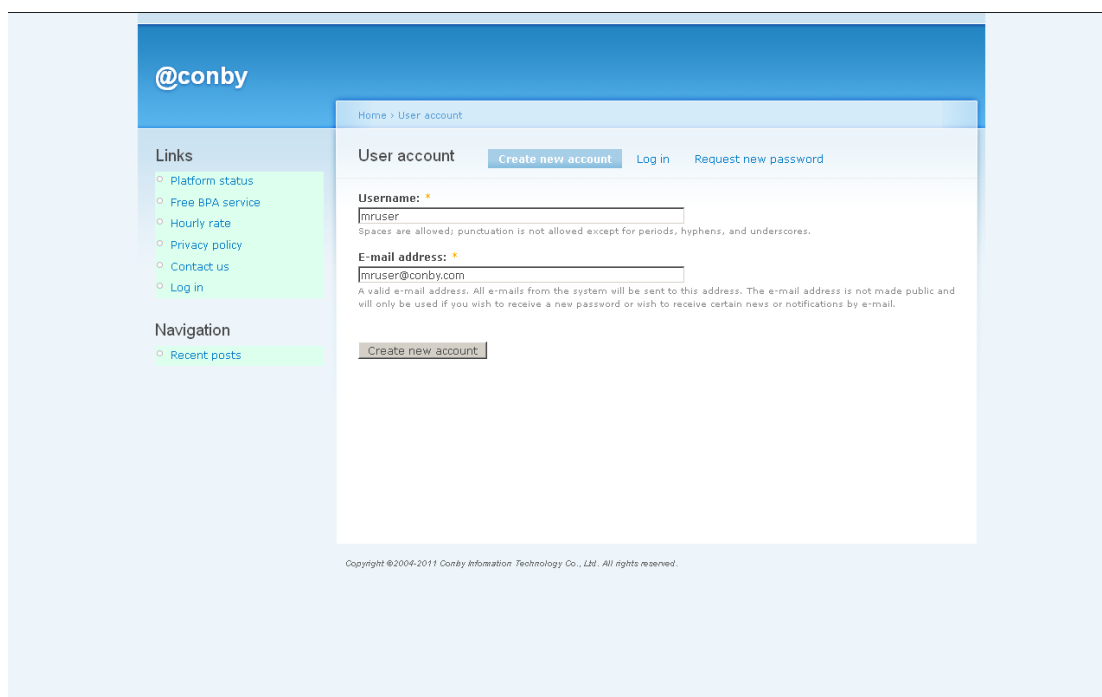
# 一、入门实例

## 1、用户注册、登录

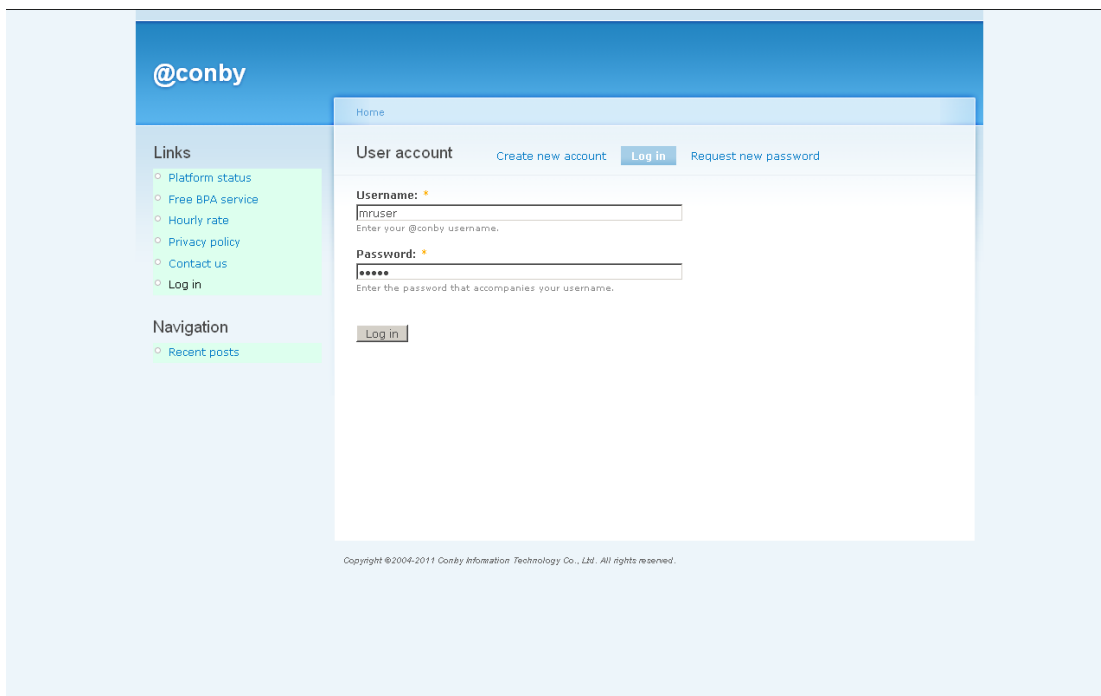
- 开始之前，请仔细阅读技术白皮书

<http://www.conby.com/download/CONBY%20C3%E8%BF%90%E8%90%A5%E5%B9%B3%E5%8F%B0%E6%8A%80%E6%9C%AF%E7%99%BD%E7%9A%AE%E4%B9%A6.v1.1.pdf>

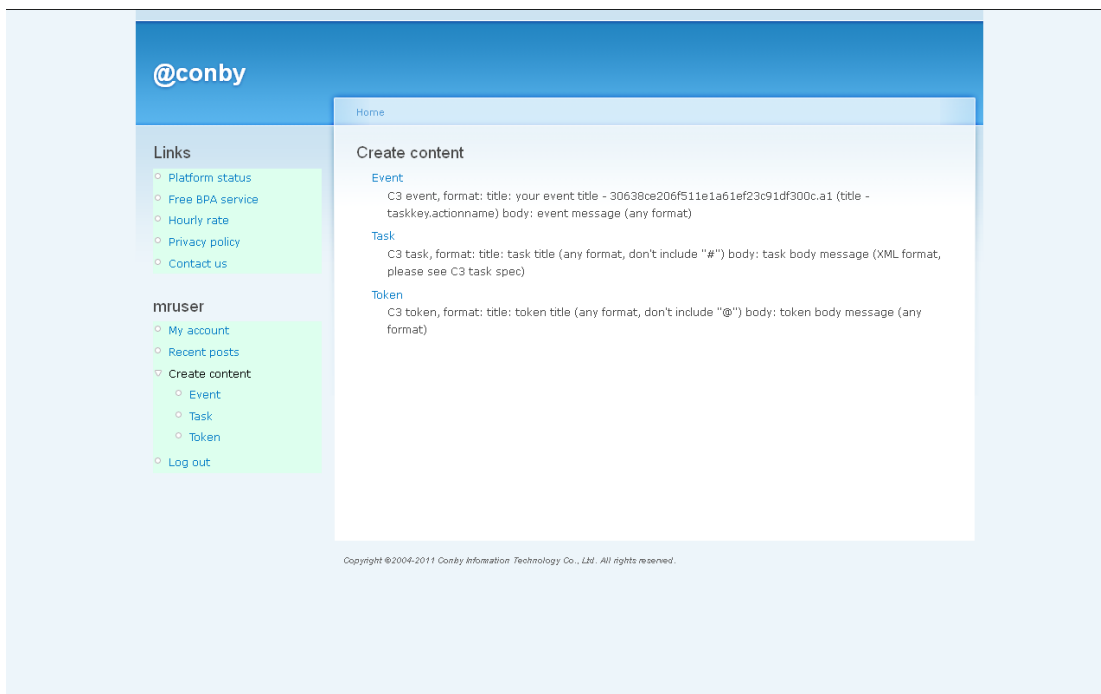
- 打开 <http://api.conby.com/user/register> 输入用户名和邮件地址，点击创建帐户，如图



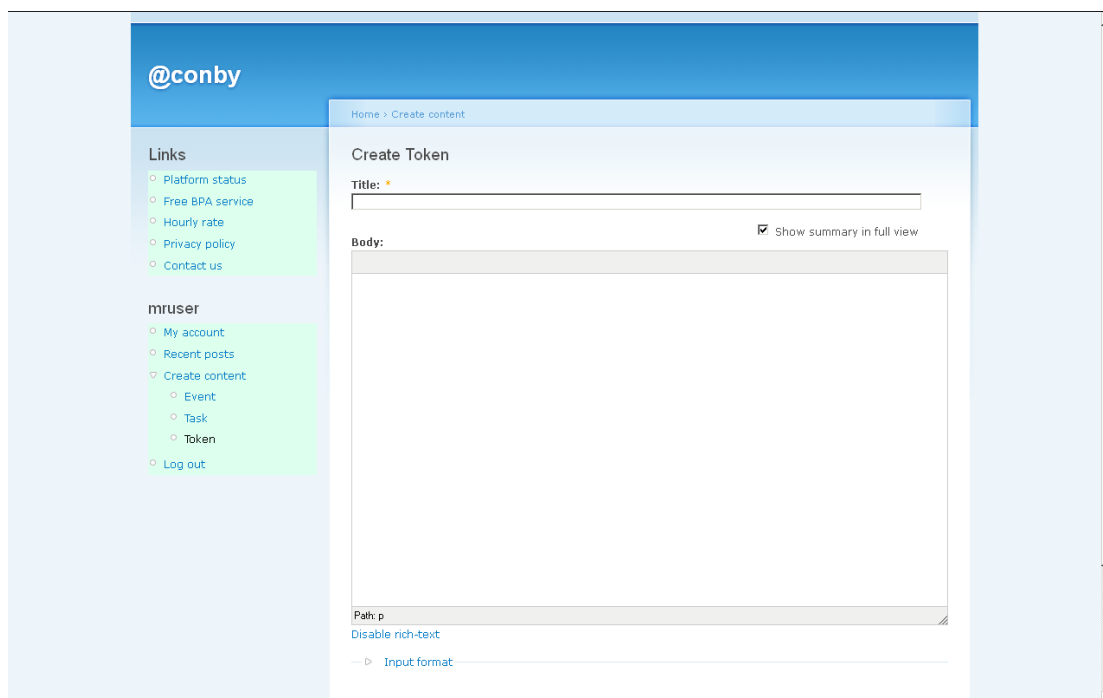
- 到你电子邮件邮箱中检查邮件，使用邮件中的用户名和密码登陆，如图



- 登录进入系统后，点击左边“Create content”进入如图界面



- 再点击 **Token** 进入创建 **Token** 的界面，至此准备工作就绪，如图



## 2、数据源搜索规则定义

- 为演示简单起见，我们使用这个网页的作为练习目标， 如果乱码，请选择 **GB2312** 编码



hello test 中文信息.abc

张三aaa李四bbb王五ccc

- 现在开始定义搜索规则，对网页进行分析后，我们可以使用如下规则

```
[default-match]
@m1=/demohtml/

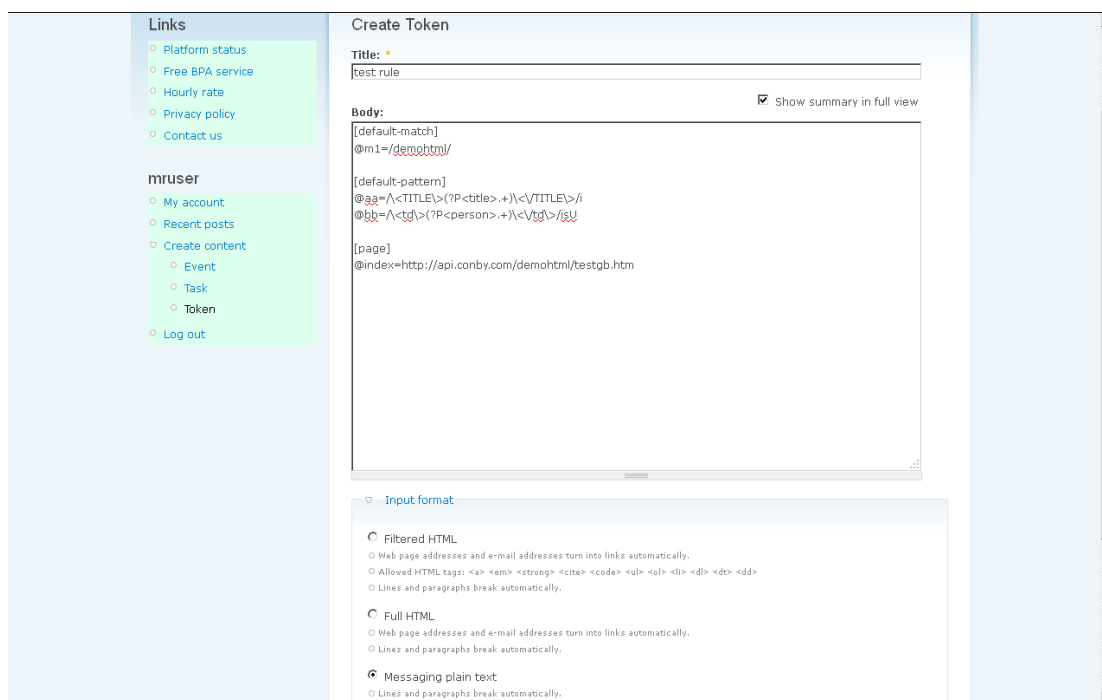
[default-pattern]
@aa=^<TITLE>(<?P<title>.+)</TITLE>/i
@bb=^<td>(<?P<person>.+)</td>/isU

[page]
@index=http://api.conby.com/demohtml/testgb.htm
```

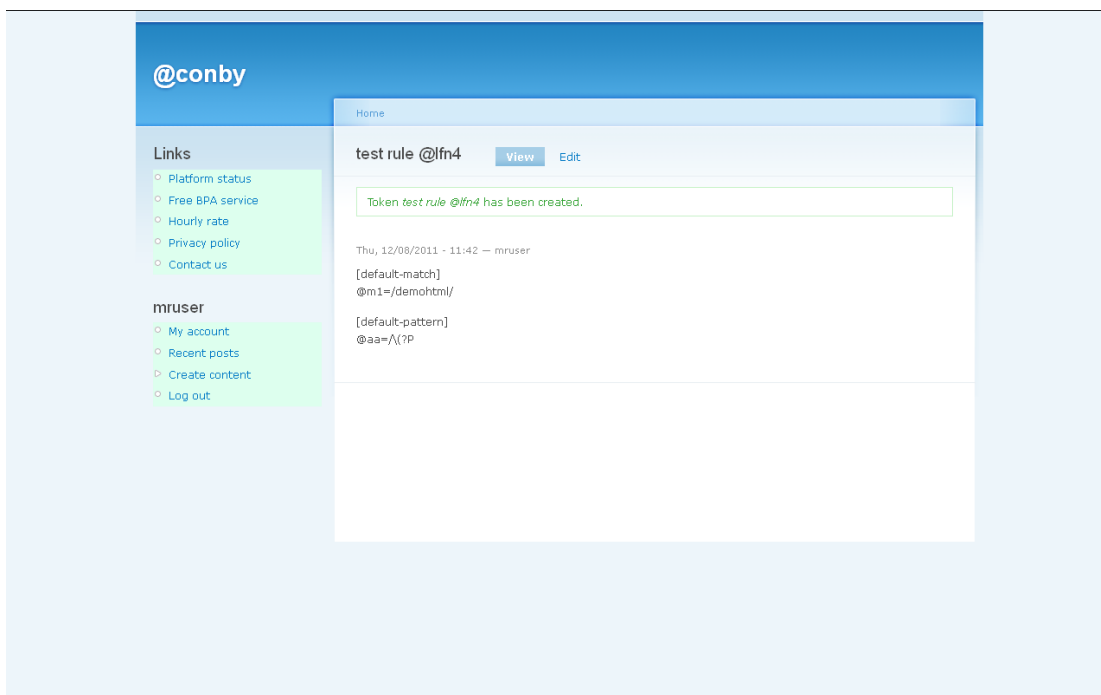
定义了一个 default 的 match 条件

同时定义了配套的数据抽取规则 default 的 pattern，我们抽取 2 种数据 title 和 person Page 中的 index 定义我们搜索的入口，因为这个网页中并没有链接，所以搜索引擎只会搜索一个页面就会完成整个搜索任务。

- 进入第一步骤中的 Create Token 界面，选择 input format 为 messaging plain text，如图



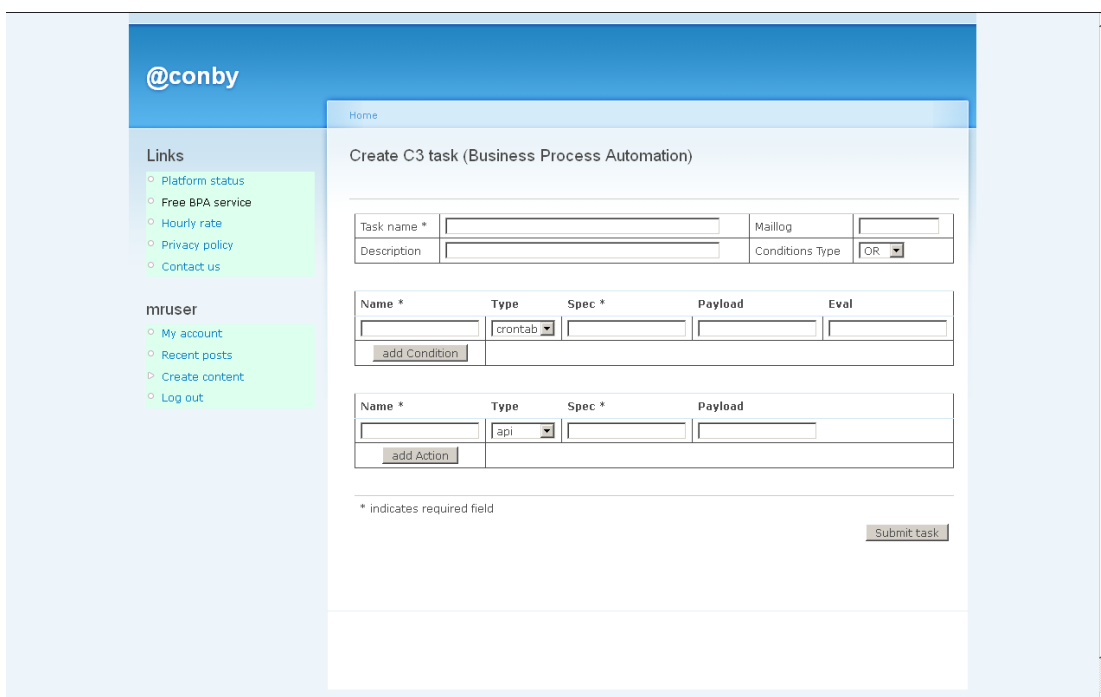
- 保存后，界面如图



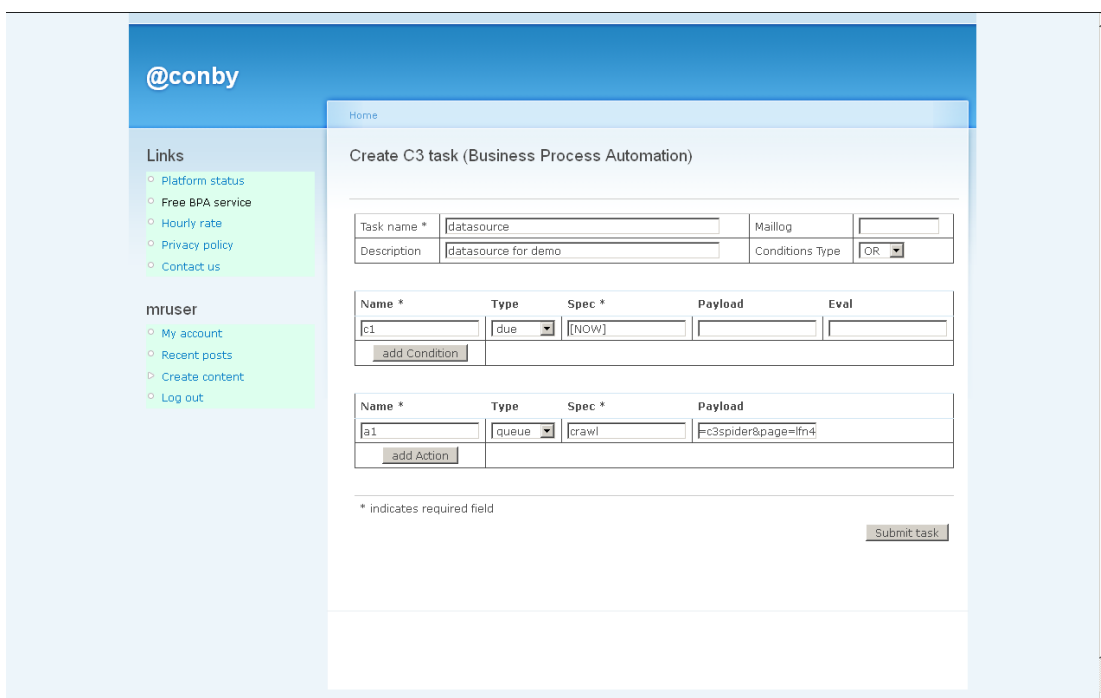
@lfn4 即为我们需要使用的 Token 名称

### 3、数据源 Task 定义

- 点击左边的“Free BPA service”链接进入创建 Task 的界面，如图



- 输入数据如图



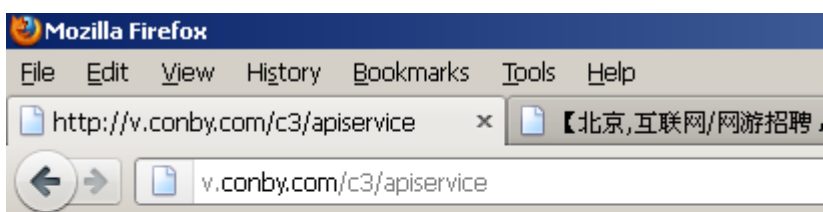
其中 payload 中输入的为:

`py=c3spider&page=lf4`

lf4 为刚才创建的 Token， 定义了搜索规则

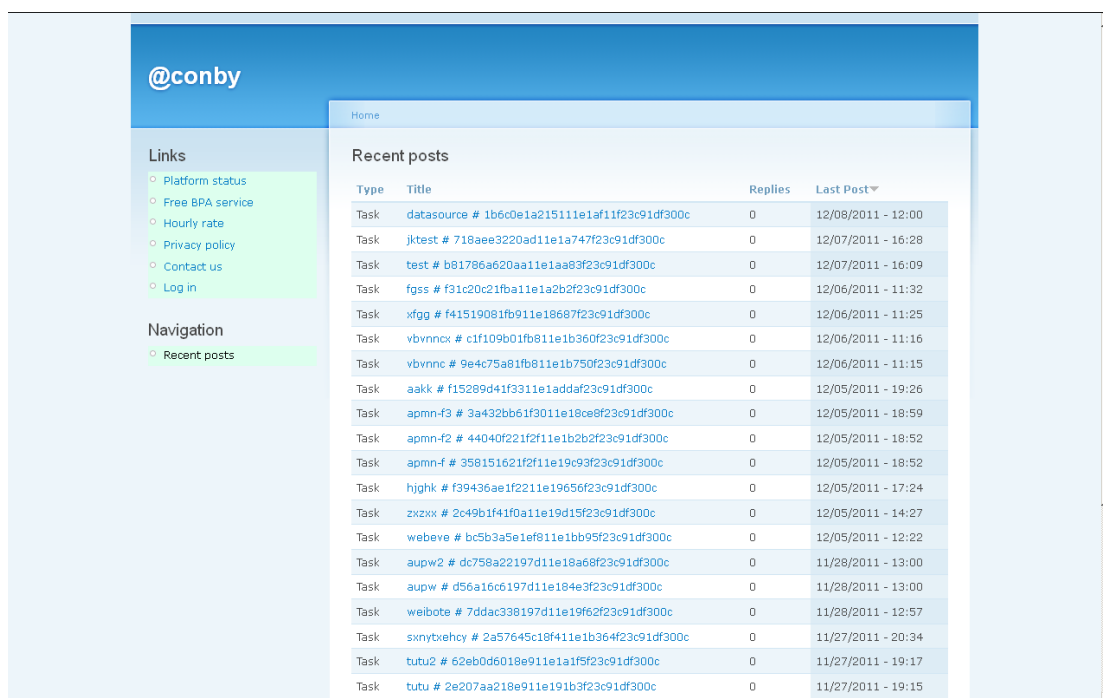
condition 中的[NOW]表示现在运行这个 Task

- 提交后界面返回 Task 的 key， 请注意保存这个 key， 可用此 key 对 task 进行管理和应用

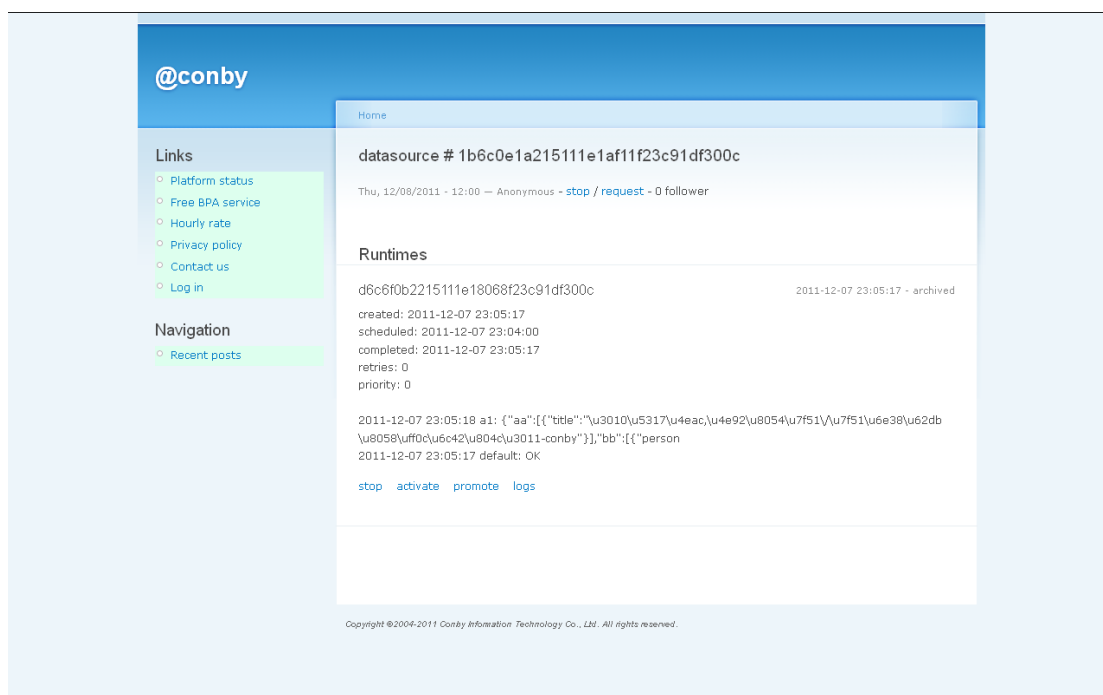


1b6c0e1a215111e1af11f23c91df300c

- 因普通用户权限原因，现在注销用户，到 Recent Post 中查看状态



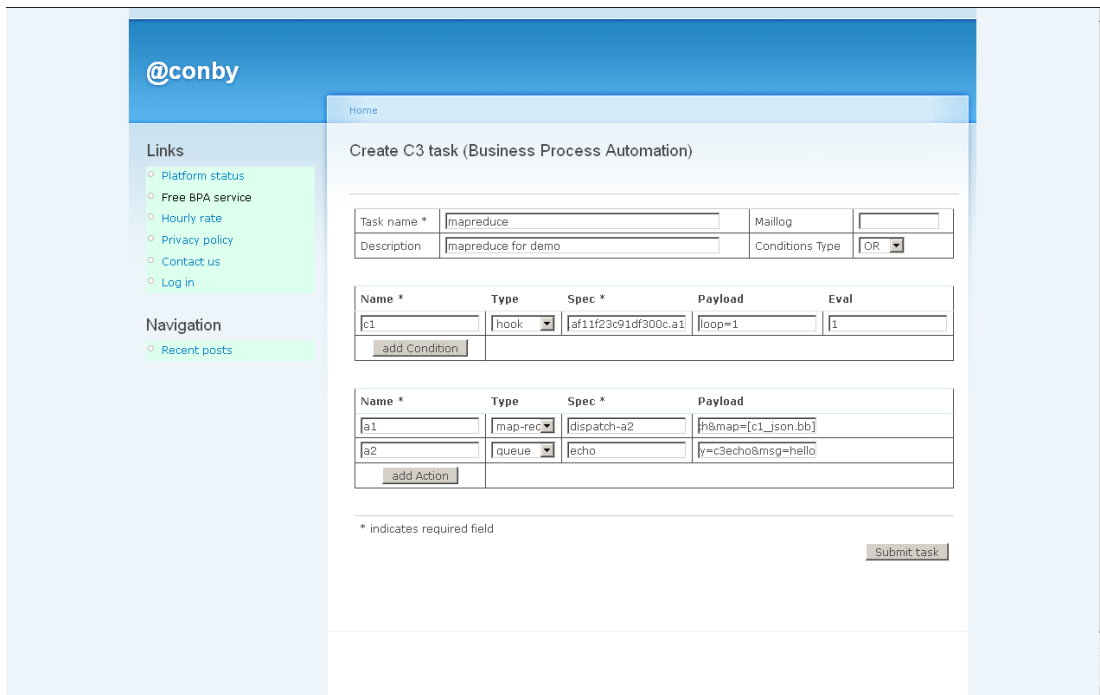
- 进入我们自己的 Task



可以观察到，Task 已经运行完毕，Action 的结果也已经显示出来

## 4、Map-Reduce 计算 Task 定义

- 点击左边的“Free BPA service”链接进入创建 Task 的界面，如图



The screenshot shows the 'Create C3 task (Business Process Automation)' interface. The form includes the following fields and tables:

Task name \*: mapreduce  
 Description: mapreduce for demo  
 Maillog:   
 Conditions Type: OR

Name *	Type	Spec *	Payload	Eval
c1	hook	af11f23e91df300c.a1	loop=1	1
<input type="button" value="add Condition"/>				

Name *	Type	Spec *	Payload
a1	map-rec	dispatch-a2	h&map=[c1_json.bb]
a2	queue	echo	y=c3echo&msg=hello
<input type="button" value="add Action"/>			

\* indicates required field

C1 condition 定义:

Type: hook

Spec: 1b6c0e1a215111e1af11f23c91df300c.a1

Payload: loop=1

Eval: 1

A1 action 定义:

Type: map-reduce

Spec: dispatch-a2

Payload: py=c3dispatch&map=[c1\_json.bb]

A2 action 定义:

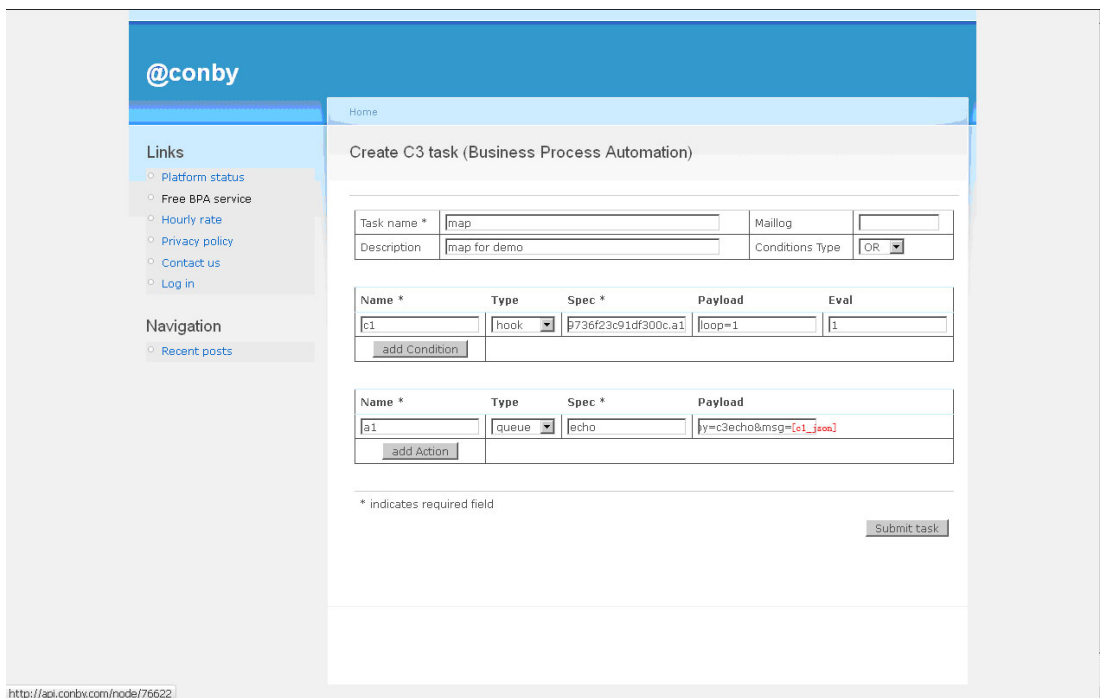
Type: queue

Spec: echo

Payload: py=c3echo&msg=hello

## 5、Map 计算 Task 定义

- 点击左边的“Free BPA service”链接进入创建 Task 的界面，如图



Home

Create C3 task (Business Process Automation)

Task name \*  Maillog

Description  Conditions Type

Name *	Type	Spec *	Payload	Eval
<input type="text" value="c1"/>	<input type="text" value="hook"/>	<input type="text" value="p736f23e91df300c.a1"/>	<input type="text" value="loop=1"/>	<input type="text" value="1"/>
<input type="button" value="add Condition"/>				

Name *	Type	Spec *	Payload
<input type="text" value="a1"/>	<input type="text" value="queue"/>	<input type="text" value="echo"/>	<input type="text" value="py=c3echo&amp;msg=[c1_json]"/>
<input type="button" value="add Action"/>			

\* indicates required field

<http://api.conby.com/node/76622>

C1 condition 定义:

Type: hook

Spec: 64a0de24215311e19736f23c91df300c.a1

Payload: loop=1

Eval: 1

A1 action 定义:

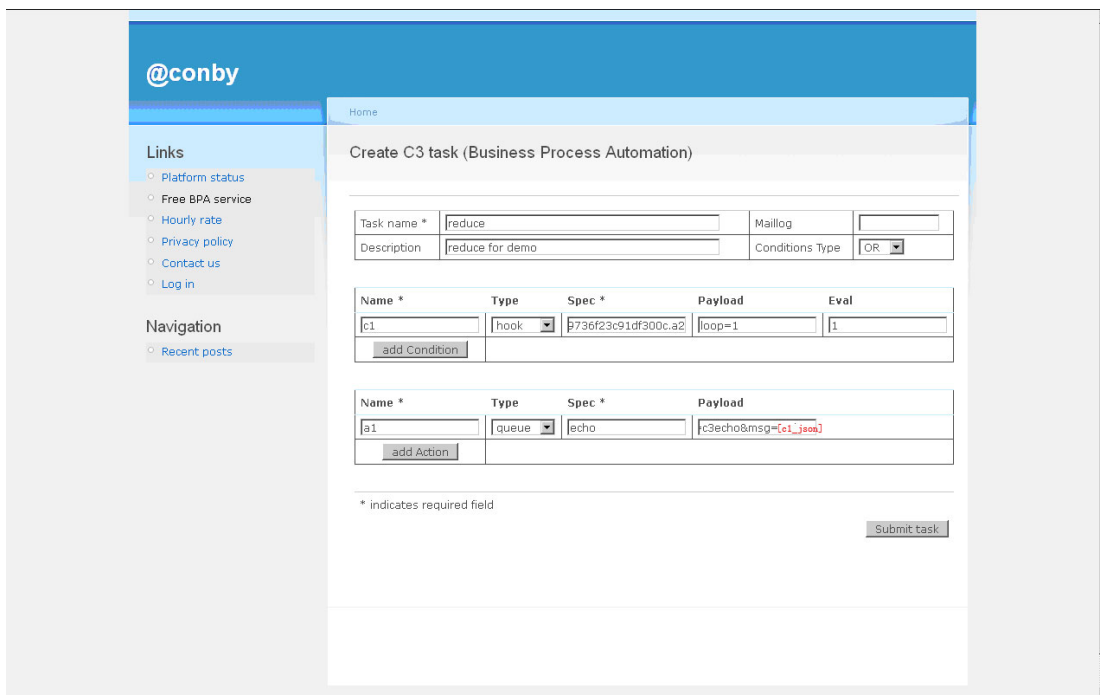
Type: queue

Spec: echo

Payload: py=c3echo&msg=[c1\_json]

## 6、Reduce 计算 Task 定义

- 点击左边的“Free BPA service”链接进入创建 Task 的界面，如图



The screenshot shows the 'Create C3 task (Business Process Automation)' interface. The form includes the following fields and tables:

Task name *	reduce	Maillog	
Description	reduce for demo	Conditions Type	OR

Name *	Type	Spec *	Payload	Eval
c1	hook	6736f23c91df300c.a2	loop=1	1

Name *	Type	Spec *	Payload
a1	queue	echo	c3echo&msg=[c1_json]

\* indicates required field

Submit task

C1 condition 定义:

Type: hook

Spec: 64a0de24215311e19736f23c91df300c.a2

Payload: loop=1

Eval: 1

A1 action 定义:

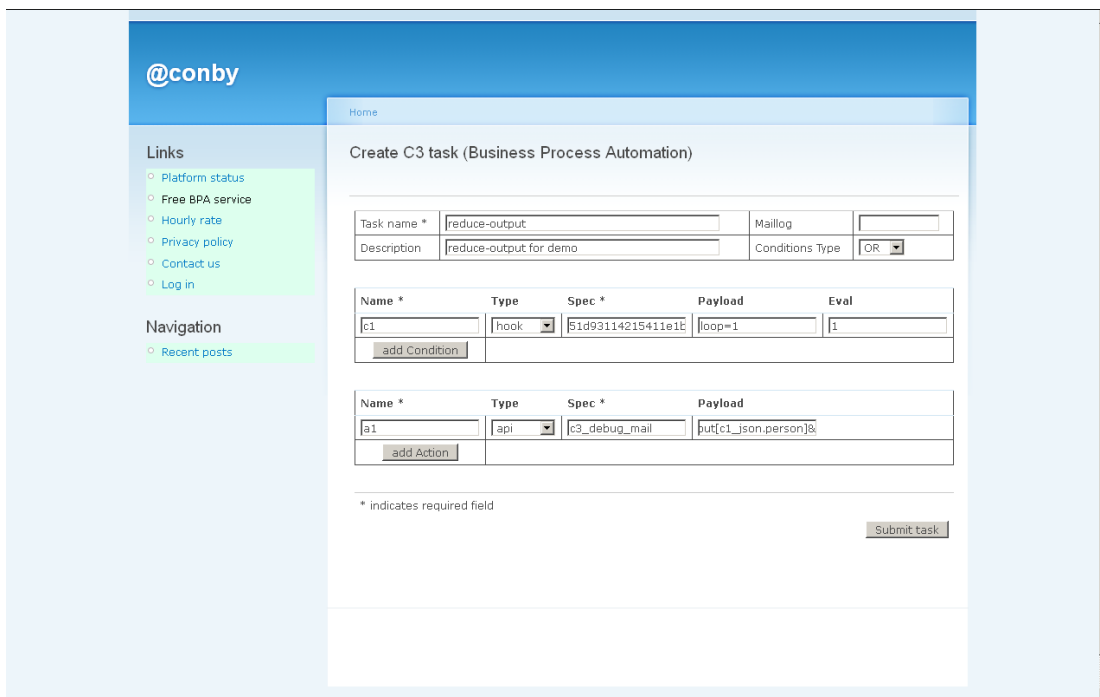
Type: queue

Spec: echo

Payload: py=c3echo&msg=[c1\_json]

## 7、Map-reduce 计算过程输出

- 点击左边的“Free BPA service”链接进入创建 Task 的界面，如图



Home

@conby

Links

- Platform status
- Free BPA service
- Hourly rate
- Privacy policy
- Contact us
- Log in

Navigation

- Recent posts

Create C3 task (Business Process Automation)

Task name \* reduce-output Maillog

Description reduce-output for demo Conditions Type OR

Name *	Type	Spec *	Payload	Eval
c1	hook	51d93114215411e1b2a7f23c91df300c.a1	loop=1	1
<input type="button" value="add Condition"/>				

Name *	Type	Spec *	Payload
a1	api	c3_debug_mail	put[c1_json.person]&
<input type="button" value="add Action"/>			

\* indicates required field

C1 condition 定义:

Type: hook

Spec: 51d93114215411e1b2a7f23c91df300c.a1

Payload: loop=1

Eval: 1

A1 action 定义:

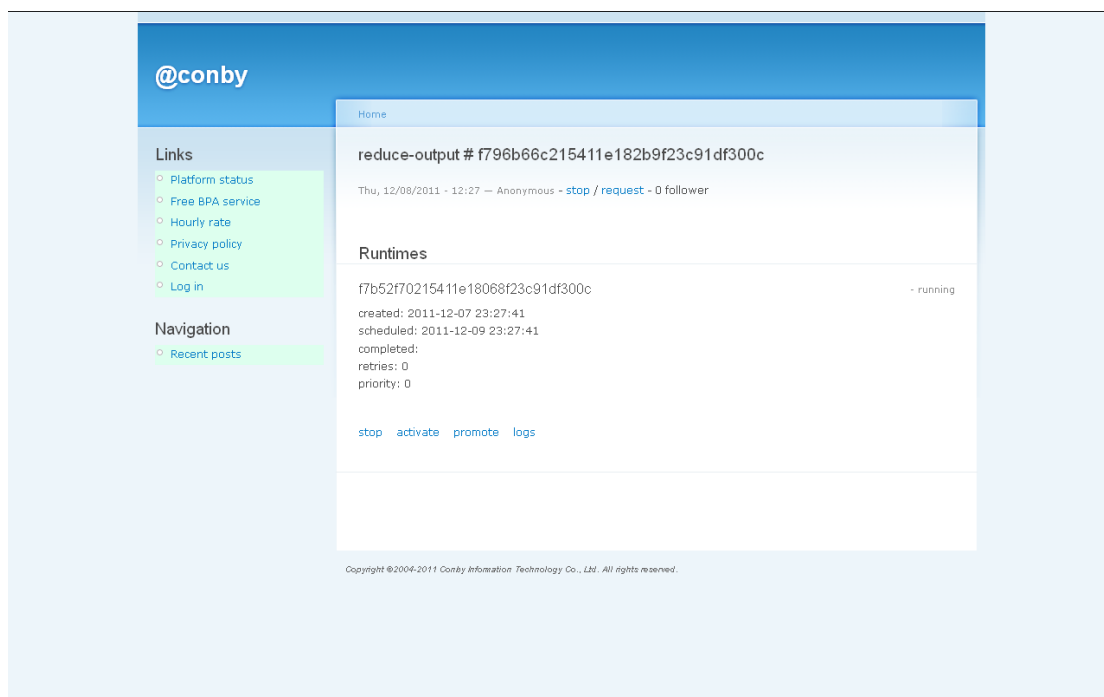
Type: api

Spec: c3\_debug\_mail

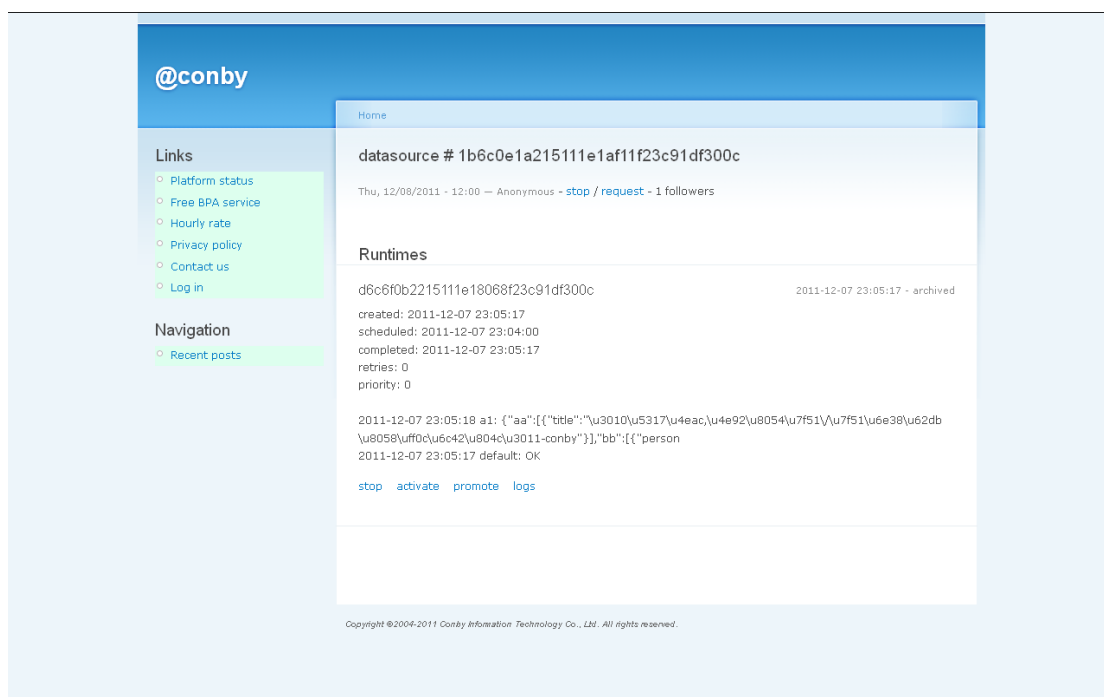
Payload: subject=reduce-output[c1\_json.person]&to=test@conby.com

可以把上面电子邮件修改为你自己的电子邮件

■ Task 状态如图



OK，现在所有流程已经设计完毕，因为数据源 Task 已经运行完毕，所以到目前为此不会产生新的数据，为了让数据源再次运行，我们可以再次激活该 Task 的 runtime，如图



点击 Runtime 下面的 activate

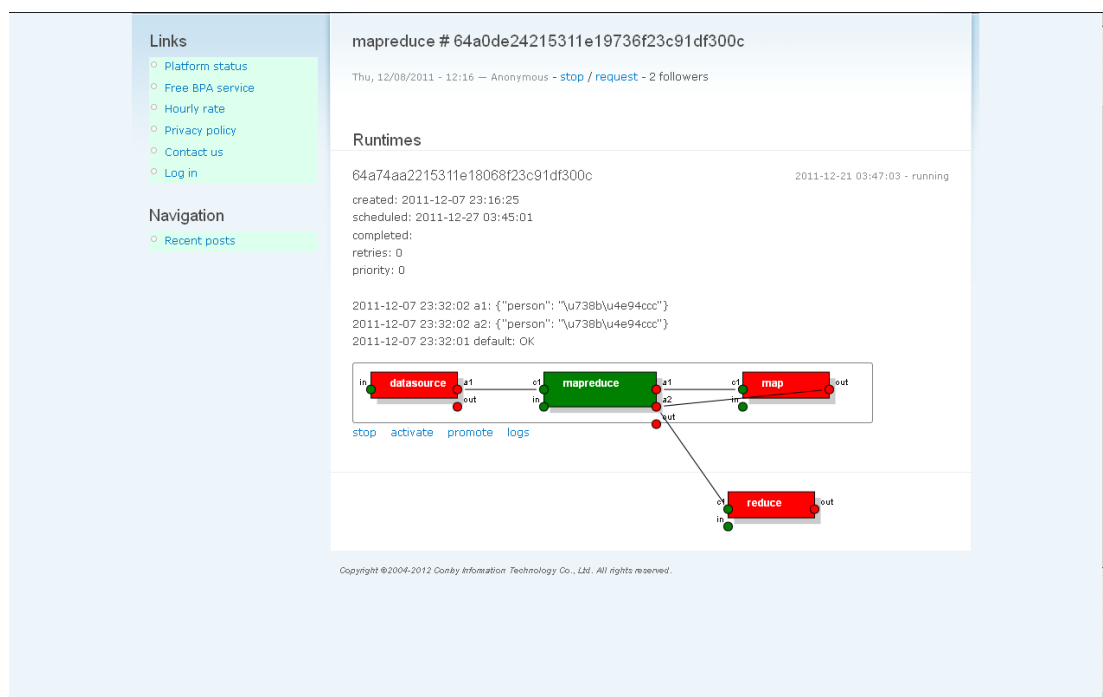
```
2011-12-07 23:05:18 a1: {"aa": [{"ti
\u8058\u0c\u6c42\u804c\u3011-coi
2011-12-07 23:05:17 default: OK
```

stop **activate** promote logs

一切正常的话，你的电子邮件里应该会收到最后运算的结果，当然我们这里 map 和 reduce 实际上什么都没做，为了演示起见，只是使用 echo 方法直接回显原来的数据，实际应用过程中需要使用自己的 API 进行业务逻辑处理。

## 8、离散事件系统流程图

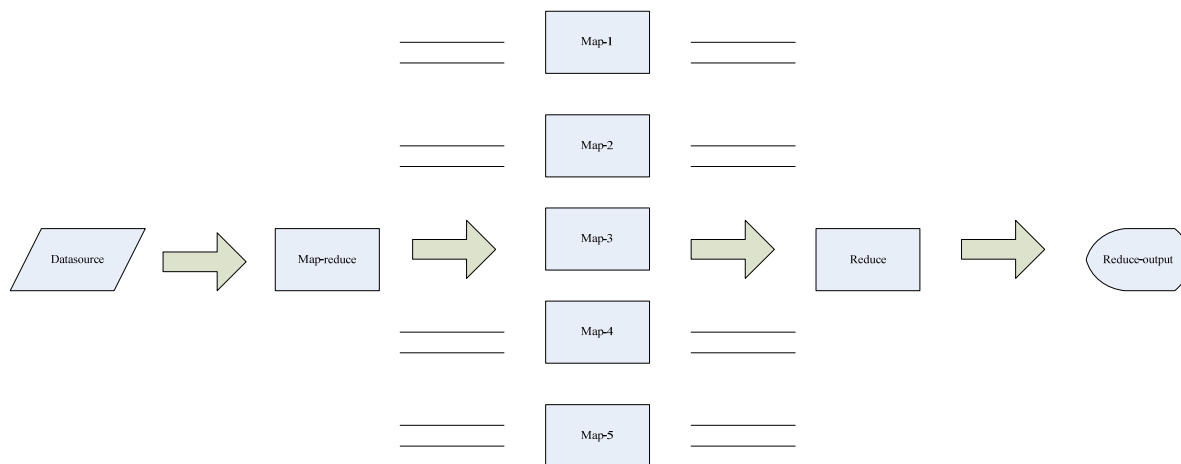
- OK，现在来看看上面模型的 DEVS(离散事件系统规范)流程图，如图



输出部分 DEVS 流程图，如图(鼠标双击 reduce 框图进入下一流程)



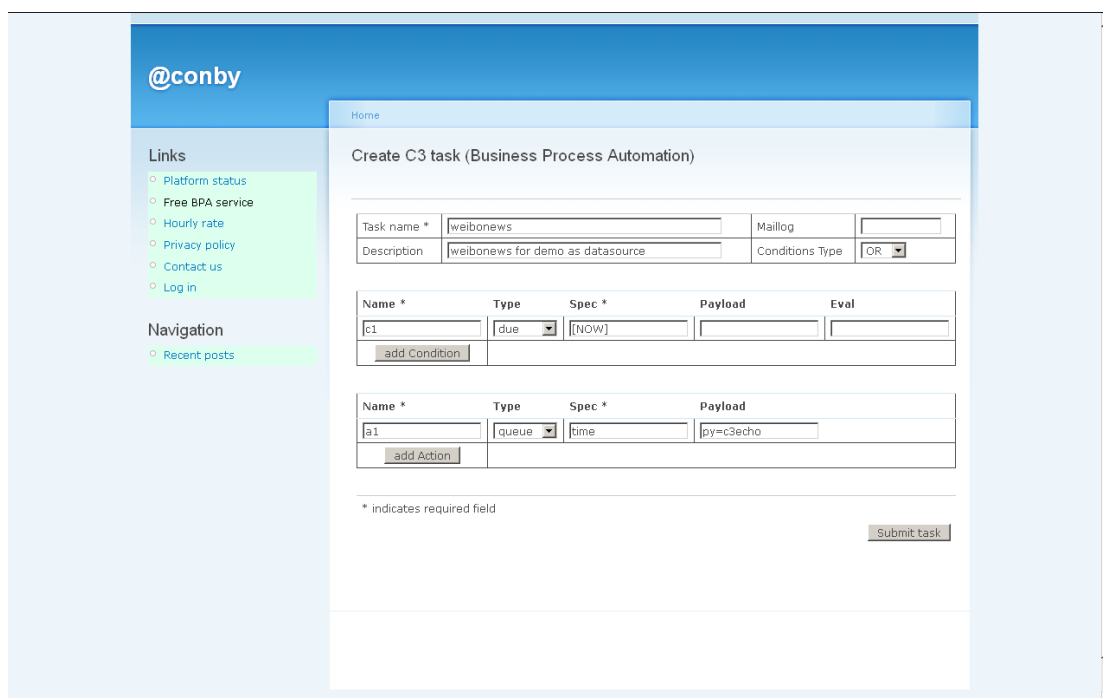
## 二、Map-Reduce 计算架构图



### 三、模拟浏览器自动登录并发微博

#### 1、微博消息源定义

- 为演示简单起见，我们使用一个简单的 Task 作为消息源



C1 condition 定义:

Type: due

Spec: [NOW]

Payload:

Eval:

A1 action 定义:

Type: queue

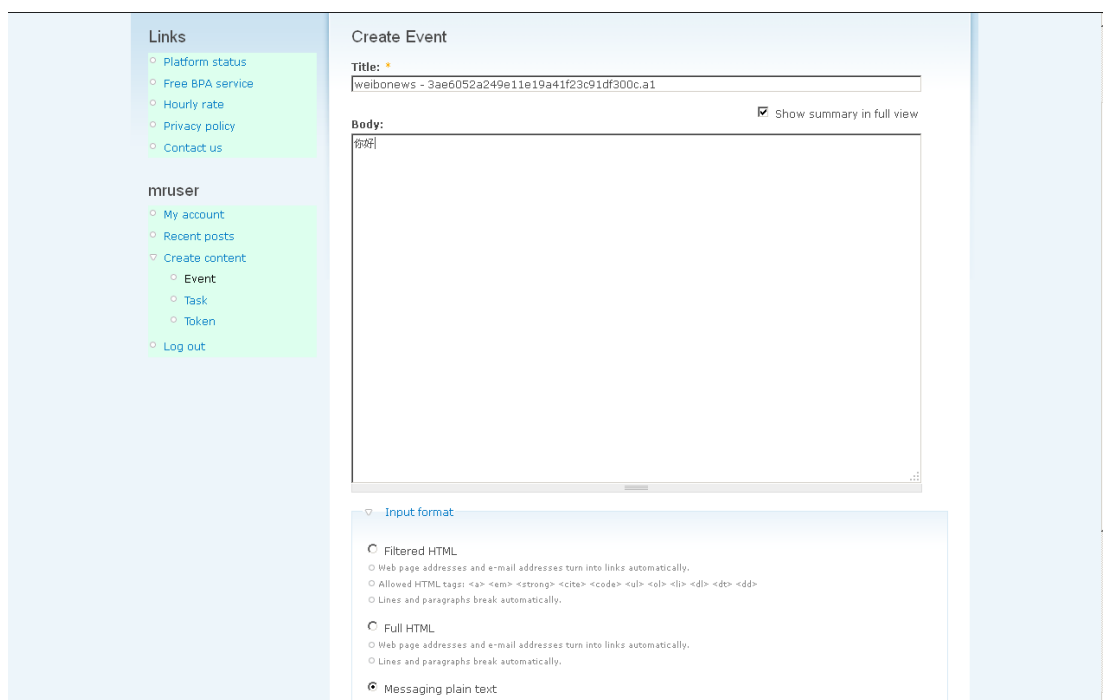
Spec: time

Payload: py=c3echo

创建好 Task, Task Key 为: 3ae6052a249e11e19a41f23c91df300c

## 2、人工微博消息事件定义定义

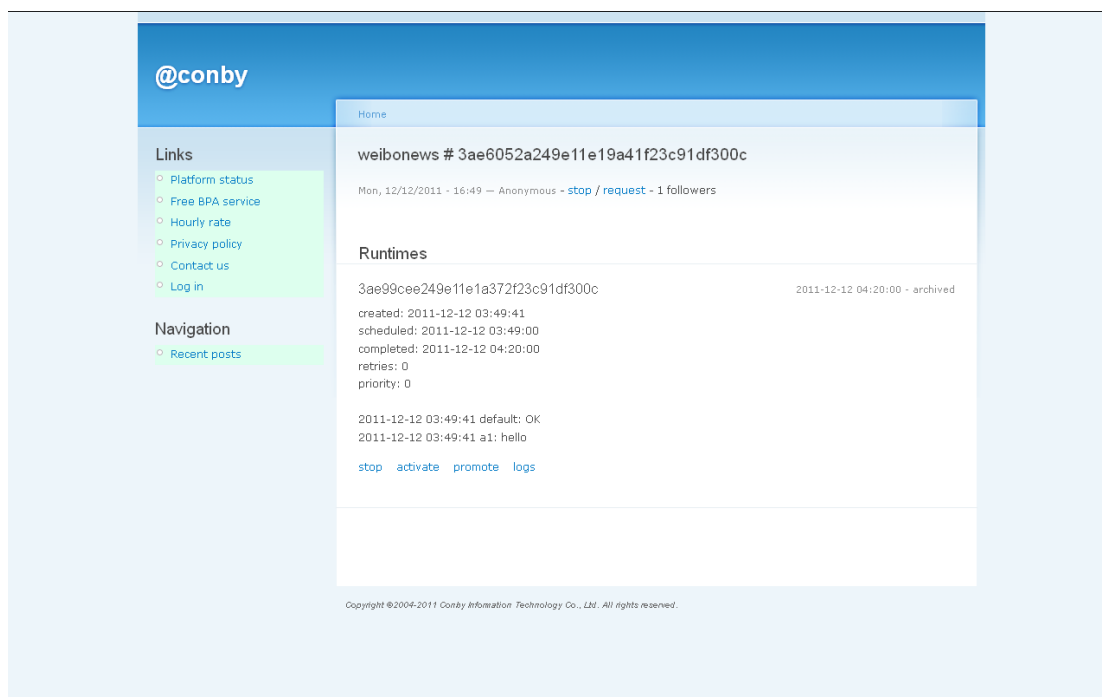
- 使用 `mruser` 用户登陆, 进入 `Create Event` 界面, 我们来创建一个人为驱动的消息事件(注意, `Input format` 格式选择 `Messaging Plain Text`)



Event Title 定义: weibonews - 3ae6052a249e11e19a41f23c91df300c.a1

Event Body 定义: 你好

观察消息源 Task, A1 的结果会受这个 body 修改结果影响, 如图修改为 hello 时:



### 3、微博登录及发送规则定义

- 现在开始定义模拟浏览器自动登录并发微博的规则，对 weibo.cn 网页进行分析后，我们可以使用如下规则

```
[default-match]
@m1=prog\wapsite\sso\login\.php
@m2=dpool\ttt\mblogDeal\.php
@m3=\.php/
@m4=crossDomain\.php

[default-pattern]
@aa=^<TITLE>(<?P<title>.+)<\</TITLE>$/

[login-form]
@url=prog\wapsite\sso\login\.php
@action=login_submit\.php
@mobile=test@conby.com
@password=conby.com
@submit=submit

[post-form]
@url=dpool\ttt\mblogDeal\.php
@action=dpool\ttt\mblogDeal\.php
@content= 天气不错 {@c1}
```

[page]

```
@index=http://3g.sina.com.cn/prog/wapsite/sso/login.php?ns=1&backTitle=title&vt=3
&backURL=http%3A%2F%2Fweibo.cn%2Fdpool%2Fttt%2FmblogDeal.php%3Fst%3
Dab8f%26act%3DshowComposer%26p%3Dtop%26vt%3D3
```

[useragent]

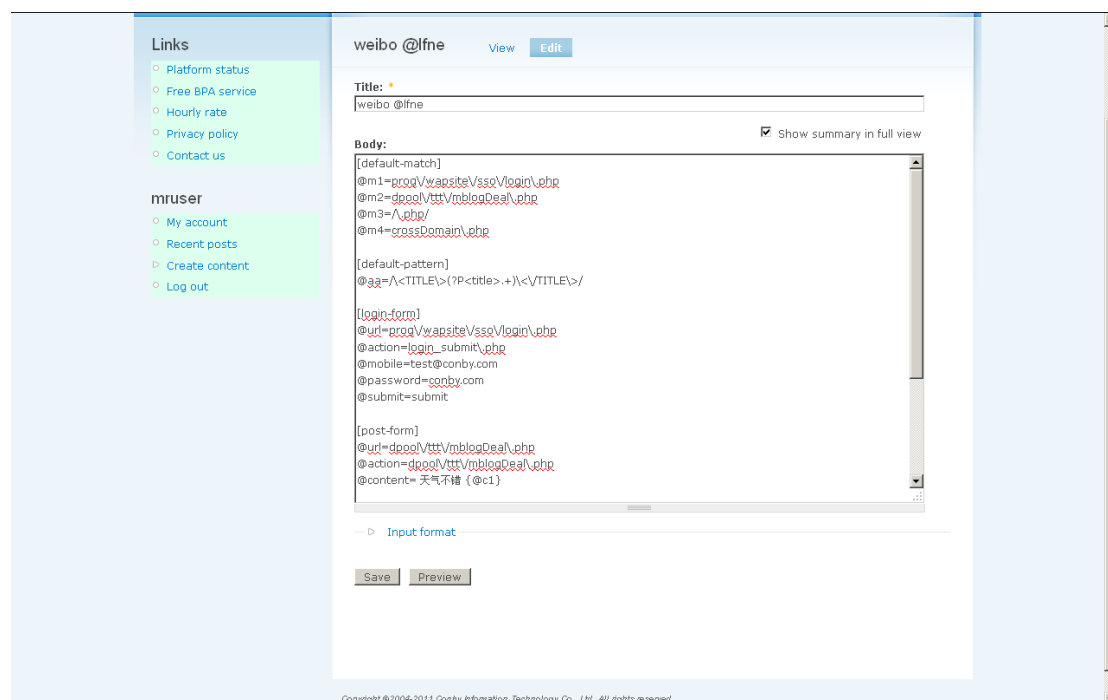
```
@ua1=Opera/9.80 (J2ME/MIDP; Opera Mini/4.0/886; U; en) Presto/2.4.15
```

定义了一个 default 的 match 条件

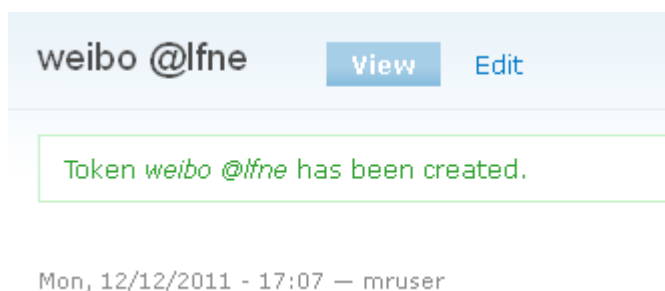
同时定义了配套的数据抽取规则 default 的 pattern，我们只抽取 2 种数据 title 作为示例，重要的部分是我们定义了两个自动 Post 数据的 Form:

- Login-form 用来模拟浏览器自动登录到新浪微博
  - Post-form 用来模拟浏览器自动发微博，我们要发送的内容是：天气不错 {@c1}，其中 {@c1} 为一个变量，用于接受 Condition c1 中的数据
- Page 中的 index 定义我们搜索的入口，为新浪微博的 SSO 登录入口。

进入 Create Token 界面,选择 input format 为 messaging plain text, 如

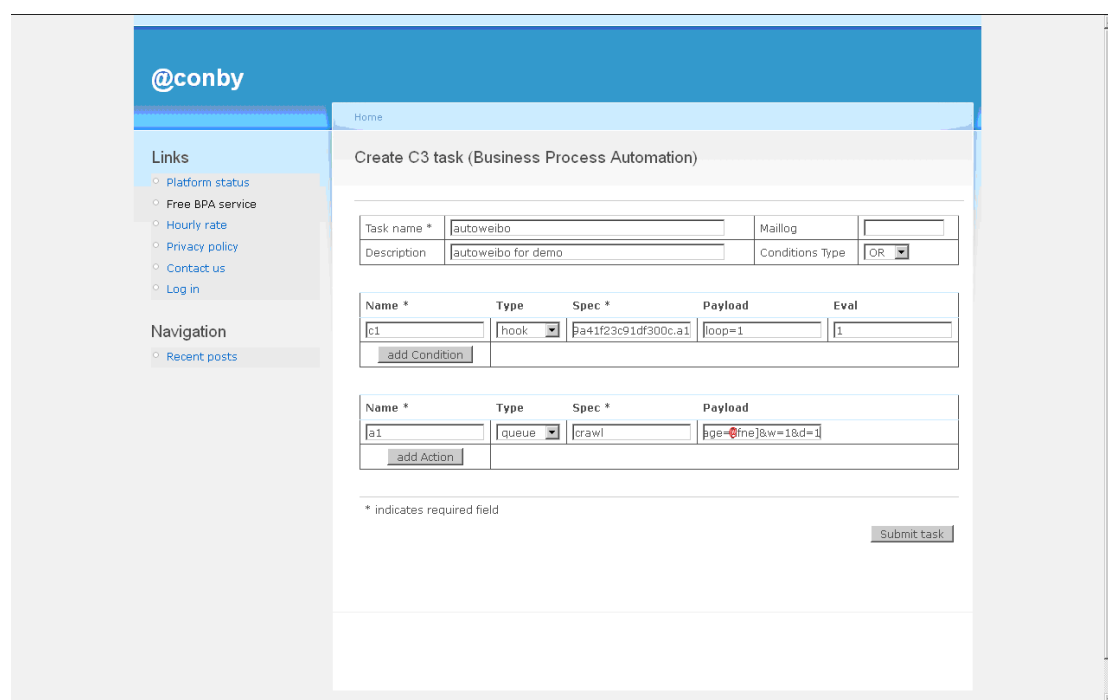


得到的 token 信息为:



## 4、微博登录及自动发微博 Task

- 进入创建 Task 的 界面，如图：



C1 condition 定义：

Type: hook

Spec: 3ae6052a249e11e19a41f23c91df300c.a1

Payload: loop=1

Eval: 1

A1 action 定义：

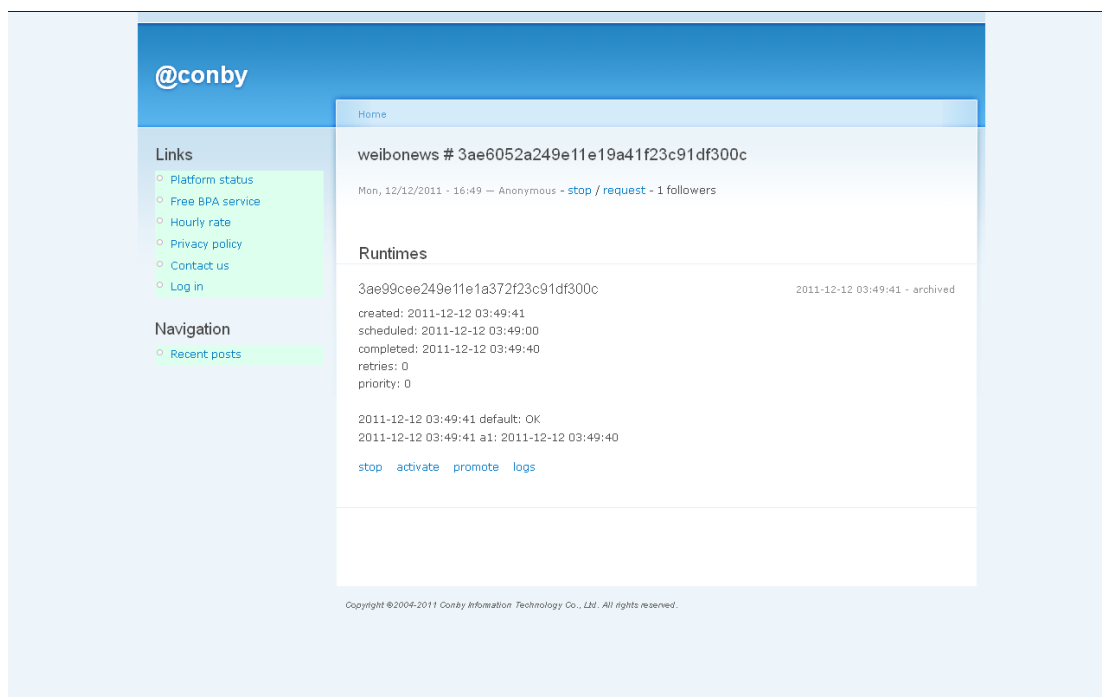
Type: queue

Spec: crawl

Payload: py=c3spider&page=[@lfne]&w=1&d=1

## 5、激活消息源触发登录及自动发微博动作

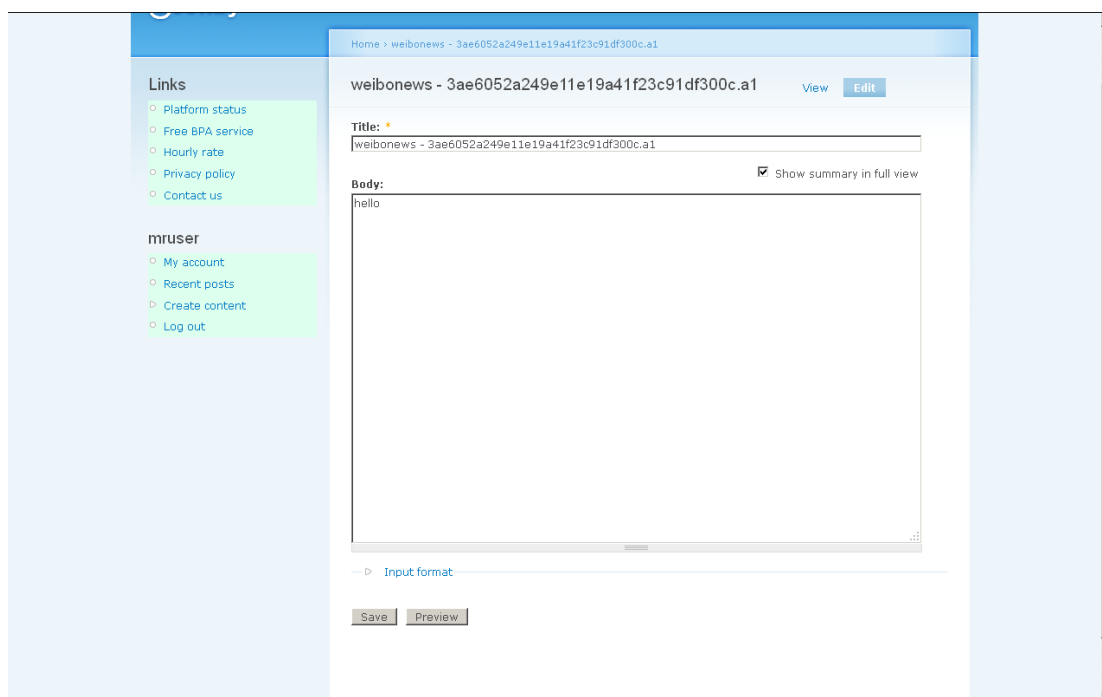
- 进入 Task 状态的 界面，如图：



- 点击 Activate，激活消息源，然后观察 Task 的状态，如图：

## 6、触发人工事件引发自动发微博动作

- 进入 Event 状态的 界面，点 Edit 进入修改状态，如图：



## 7、检查微博验证结果

- 进入weibo.com, 使用 [test@conby.com](mailto:test@conby.com) 登录, 结果如图:



因为我们使用了手机浏览器的 User Agent,所以新浪微博显示消息来自于 手机版。

为防止服务器 IP 被网站封掉, 一般情况下必须使用代理服务器进行测试, 如下:

```
[proxy]
@proxy1=http://208.208.208.208:8080/
```

注意: 请使用有效的代理服务器地址

## 8、离散事件系统流程图

现在看 DEVS(离散事件系统规范)流程图, 结果如图:

```
2011-12-12 04:31:15 default: OK
2011-12-12 04:31:15 a1: {"c3_ex":{"href":[],"event":[]}}
```

