# Website Input - Feature #2220

## Stream results without caching them to reduce memory usage

02/24/2018 07:49 PM - Luke Murphey

| | | | |
|---|---|---|---|
| **Status:** | Closed | **Start date:** | 02/24/2018 |
| **Priority:** | Normal | **Due date:** | |
| **Assignee:** | Luke Murphey | **% Done:** | 100% |
| **Category:** | Input: Web Spider | **Estimated time:** | 0.00 hour |
| **Target version:** | 4.5.2 | | |

**Description**

### Associated revisions

**Revision 644 - 03/02/2018 05:53 AM - lukemurphey**

Adding test server that includes a large chunk of text

This is useful for performance testing

Reference #2220

**Revision 645 - 03/02/2018 05:53 AM - lukemurphey**

Adding streaming of results as they arrive

Reference #2220

**Revision 647 - 03/02/2018 09:26 PM - lukemurphey**

Increasing the size of test output to increase chance of seeing performance problems

Reference #2220

**Revision 648 - 03/02/2018 09:45 PM - lukemurphey**

Forcing garbage collection to make performance analysis easier

This also reduce overall memory usage by cleaning some things up
while the input waits for the next run

Reference #2220

### History

**#1 - 02/24/2018 08:07 PM - Luke Murphey**

**web_input.py:** calls output_event in the modular input class

To make this work, I would need to:

1. Pass an output result function to **scrape_page()**
2. Pass the output result function to **get_result_single()**

3. Keep a result count around. Maybe just having **get_result_single()** return a small set of fields, like just the URL.

#### #2 - 02/25/2018 08:27 PM - Luke Murphey

Before I do the change I should test this and monitor memory usage. I could do this by patching the internal web-server to return content with variable URLs that should cause the existing design to load up memory.

Then, I can run this afterwards to show that memory usage is down.

#### #3 - 02/25/2018 08:44 PM - Luke Murphey

*- Target version changed from 4.5.3 to 4.5.2*

#### #4 - 02/27/2018 02:25 AM - Luke Murphey

Monitoring memory with:

```
source=top PID=57782 | timechart max(RES) as memory
```

#### #5 - 03/02/2018 12:39 AM - Luke Murphey

It seems like some memory isn't being reclaimed. When I use output only when contents change, the memory usage seems like it is actually lower and reclaims faster once the input is done.

Next steps:

1. Disable streaming, see if the problem changes
2. Disable parts of the output function, see if the problem changes

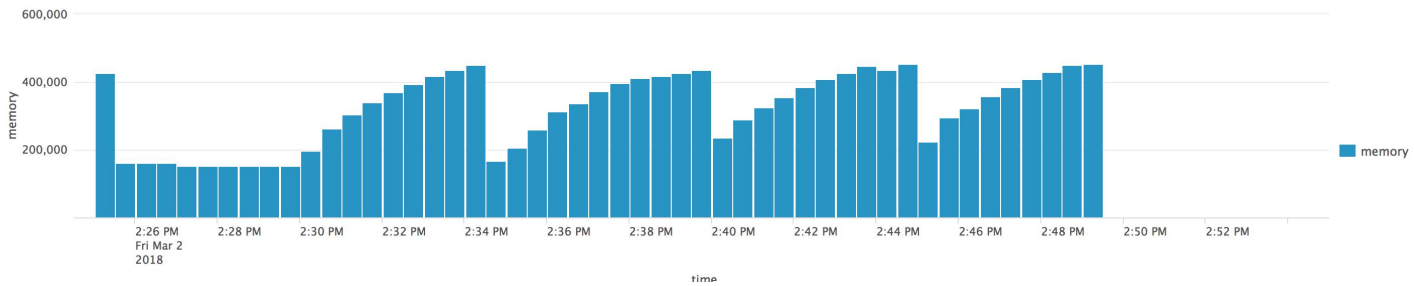
#### #6 - 03/02/2018 12:39 AM - Luke Murphey

*- % Done changed from 0 to 70*

#### #7 - 03/02/2018 09:09 PM - Luke Murphey

*- File non-streaming.png added*

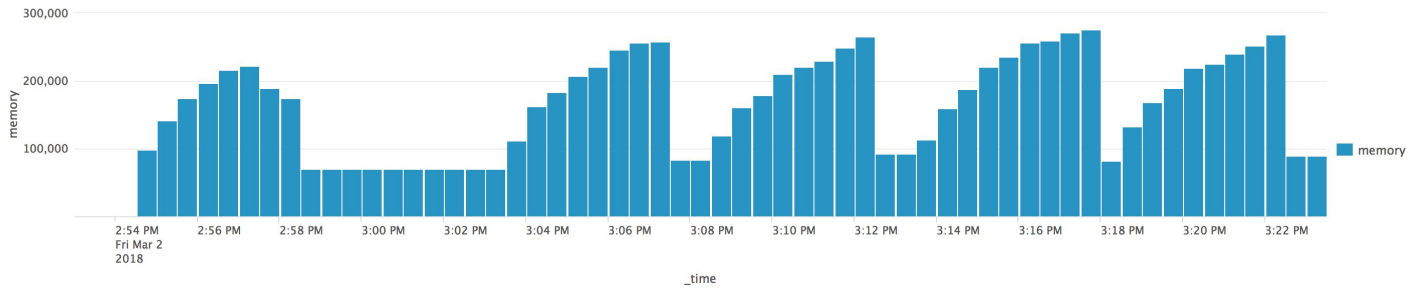
Without streaming:



**#8 - 03/02/2018 09:23 PM - Luke Murphey**

*- File streaming.png added*

With streaming:




**#9 - 03/02/2018 10:27 PM - Luke Murphey**

*- Status changed from New to Closed*

*- % Done changed from 70 to 100*

### Files

| | | | |
|---|---|---|---|
| non-streaming.png | 88.5 KB | 03/02/2018 | Luke Murphey |
| streaming.png | 93.8 KB | 03/02/2018 | Luke Murphey |