Website Input - Feature #1882

Restrict inputs to HTTPS sites if on cloud

05/25/2017 07:41 PM - Luke Murphey

Status:	Closed	**Start date:**	05/25/2017
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:	Input: Web Spider	**Estimated time:**	0.00 hour
Target version:	4.3.0		

Description

Associated revisions

Revision 444 - 07/07/2017 09:27 PM - lukemurphey

Restricting access on Splunk Cloud to HTTPS

Reference #1882

Revision 446 - 07/08/2017 05:45 PM - luke.murphey

Changes examples to use HTTPS

Reference #1882

Revision 447 - 07/08/2017 06:11 PM - luke.murphey

Making icon call use HTTPS

Reference #1882

Revision 448 - 07/08/2017 07:06 PM - lukemurphey

Making the web_scrape command ensure that connections use HTTPS on Cloud

Reference #1882

Revision 450 - 07/09/2017 04:59 AM - lukemurphey

Making sure link extraction requires HTTPS on Cloud

Reference #1882

History

#1 - 07/07/2017 08:57 PM - Luke Murphey

- Status changed from New to In Progress

- Assignee set to Luke Murphey

#2 - 07/07/2017 09:15 PM - Luke Murphey

To update:

- **[done]** Mod input editor
- **[done]** Modular input code
- **[done]** Wizard view
- **[done]** Preview controller
- **[done]** Search command
- **[done]** Spider link extraction

#3 - 07/08/2017 08:13 AM - Luke Murphey

From some reason the freaking endpoint isn't showing up in SplunkWeb.

Observations:

- Removing the other entries from web.conf doesn't help
- The entry looks equialvent to the one used in Website Montprinh
- http://127.0.0.1:8000/en-US/splunkd/services/admin/app_website_input/default fails
- https://127.0.0.1:8090/services/admin/app_website_input works
- Is not included in C:\Program Files\Splunk\var\run\splunk\merged

#4 - 07/08/2017 08:21 AM - Luke Murphey

It is available under http://127.0.0.1:8000/en-US/splunkd/__raw/services/admin/app_website_input/default

#5 - 07/08/2017 08:32 AM - Luke Murphey

- % Done changed from 0 to 80

#6 - 07/08/2017 05:10 PM - Luke Murphey

Wierd, the non __raw endpoint works on Mac and 6.6.0.

#7 - 07/08/2017 06:58 PM - Luke Murphey

I set the form value of the URL with the following to force the call to attempt to scrape the page:

```
document.getElementById('inputURL').value = "http://textcritical.net"
```

#8 - 07/09/2017 04:57 AM - Luke Murphey

I have to update several calls to pass the https_only parameter to:

- scrape_page
- get_result_single
- extract_links

#9 - 07/09/2017 04:58 AM - Luke Murphey

Tested with:

```
| webscrape selector="h3" url="https://www.reddit.com/r/popular/" page_limit=50 url_filter="*" depth_limit=25
empty_matches=0
```

#10 - 07/10/2017 01:33 AM - Luke Murphey

I want to make tests for this.

To do this, I need:

- decorator for run_only_on_cloud
- decorator for run_only_on_enterprise
- Tests for Cloud:
  - Scrape page doesn't extract non-HTTPS links
  - Controller doesn't extract non-HTTPS links
  - Scrape page won't scan non-HTTPS links
  - Controller won't scan non-HTTPS links
  - Wizard: rejects non-HTTPS

#11 - 07/10/2017 03:04 AM - Luke Murphey

I wonder if I should change scrape_page to throw an exception if the URL provided is not HTTPS. Currently, https_only is only applicable to link extractions.

#12 - 07/10/2017 03:26 AM - Luke Murphey

I recall now why I didn't add this into scrape_page: it doesn't have the session key to lookup whether the host is in the Cloud.

#13 - 07/10/2017 03:37 AM - Luke Murphey

- Status changed from In Progress to Closed

- % Done changed from 80 to 100