Parsing Apache Logs with Splunk for real user hits....

This little article is trying to find the "real user" browsing hits on my Homepage. In front of the homepage is an Apache2 acting as proxy.

All logs are gathered in Splunk. In the next part, "crawlers" is a synomyn for Robots, Crawlers and Spiders - hence non-live machines.

A log sample

```
77.243.52.139 - - [30/Oct/2017:19:55:53 +0100] "GET /display/public/HealthCheckPage HTTP/1.1" 200 13568 "-"
"Datadog Agent/5.12.3"
216.244.66.237 - - [30/Oct/2017:19:55:54 +0100] "GET /label
/alfresco+anette+apache+atlassian+christopher+cms+confluence+cookbook+esdh+groovy+itil+itsm+javascript+jira+jira
-development+jira-workflow+noshow+scriptrunner-plugin+utf-8 HTTP/1.1" 200 57729 "-" "Mozilla/5.0 (compatible;
DotBot/1.1; http://www.opensiteexplorer.org/dotbot, help@moz.com)"
216.244.66.237 - - [30/Oct/2017:19:55:55 +0100] "GET /label
/alfresco+anette+apache+atlassian+christopher+cms+confluence+cookbook+esdh+groovy+itil+itsm+javascript+jira+jira
-development+logrotate+noshow+toke HTTP/1.1" 200 57273 "-" "Mozilla/5.0 (compatible; DotBot/1.1; http://www.
opensiteexplorer.org/dotbot, help@moz.com)"
54.243.143.134 - - [30/Oct/2017:19:54:13 +0100] "HEAD /display/ATLASSIAN/Mail+workaround+for+private+setup HTTP
/1.1" 200 598 "-" "MBCrawler/1.0 (https://monitorbacklinks.com)"
162.119.128.141 - - [30/Oct/2017:19:57:04 +0100] "GET /rest/quickreload/latest/67764273?
since=1509389795941&_=1509389820540 HTTP/1.1" 204 200 "http://www.mos-eisley.dk/display/it/Beats+for+splunk"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100
Safari/537.36"
157.55.39.132 - - [30/Oct/2017:19:49:18 +0100] "GET /display/ITSM/IT+Service+Management+systemer HTTP/1.1" 200
15609 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"
85.191.200.41 - - [30/Oct/2017:19:58:10 +0100] "POST /rest/analytics/1.0/publish/bulk HTTP/1.1" 200 378
"http://www.mos-eisley.dk/pages/viewpage.action?pageId=85033157" "Mozilla/5.0 (Macintosh; Intel Mac OS X
10_13_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36"
85.191.200.41 - - [30/Oct/2017:19:58:14 +0100] "GET /pages/editpage.action?pageId=85033157 HTTP/1.1" 200 19238
"http://www.mos-eisley.dk/pages/viewpage.action?pageId=85033157" "Mozilla/5.0 (Macintosh; Intel Mac OS X
10_13_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36"
85.191.200.41 - - [30/Oct/2017:19:54:57 +0100] "GET /pages/viewpage.action?pageId=85033157 HTTP/1.1" 200 17835
"http://www.mos-eisley.dk/pages/resumedraft.action?draftId=85033158&draftShareId=3d2e60a7-ec5b-4488-8e40-
fa4d0f5a0e8d" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.
0.3163.100 Safari/537.36"
```

No valid Data Center license found
Please go to **Atlassian Marketplace** to purchase or evaluate Refined Toolkit for Confluence Data Center.
Please read this **document** to get more information about the newly released Data Center version.

Step - Filter page views

First, we need to **include** what actually is a "view" and not REST, saving pages, uploading stuff etc. For Confluence, all views are in one of 2 forms:

```
(uri="*/display/*" OR uri="*/viewpage.action/*")
```

All other URI's are not relevant...

Step - Eliminate all "bots"

Looking into the log files, look at the User Agent string, often these have a Bot-like name, but nowadays many crawlers acts as a normal browser and are not identifiable via the User Agent.

So, we try to eliminate them with **exclusions**:

```
useragent!="*bot*" useragent!="*spider*" useragent!="*facebookexternalhit*" useragent!="*crawler*" useragent!="
*Datadog Agent*"
```

Step - Eliminate Monitoring

Monitoring Tools monitoring can fill a lot in the logs; to control and identify these, I have ensured the monitoring tool is only monitoring at a special URL: /display/public/HealthCheckPage.

Hence, to **exclude** the monitring:

```
uri!="/display/public/HealthCheckPage"
```

Step - Eliminate hosts that has looked at robots.txt

To remove hits from IP-Addresses that have looked at robots.txt, I have created a lookup to a csv file.

So a scheduled Report is running hourly:

```
index=apache robots.txt clientip="*" | table clientip
```

Stored in the file robots_spiders.csv

```
root@splunkserver:/splunk/etc/apps/moseisleymonitoring/lookups# head robot_spiders.csv
clientip
"216.244.66.237"
"77.75.76.163"
"77.75.77.62"
"216.244.66.237"
"77.75.78.162"
"216.244.66.237"
"77.75.76.165"
"37.9.113.190"
"106.120.173.75"
```

To **exclude** these IP Addresses:

```
NOT [| inputlookup robot_spiders.csv | fields clientip]
```

Step - Eliminate all "hard hitting hosts"

As many crawlers use browser like User Agents and acts like real browsers, looking into my logs I see a large number of hits from them, so I have taken the assumption that more than 100 hits on the same URI within 30 days states that it is not a person using a browser.

So a scheduled Report is running daily:

```
index=apache AND host=moserver AND (uri="*/display/*" OR uri="*/viewpage.action/*") | stats count by uri
clientip | where count>100
```

Stored in the file hard_hitting_hosts.csv

```
root@splunkserver:/splunk/etc/apps/moseisleymonitoring/lookups# head hard_hitting_hosts.csv
uri,clientip,count
"/display/ATLASSIAN/JIRA+as+CMDB/","188.163.74.19",125
"/display/ATLASSIAN/JIRA+as+CMDB/","37.115.189.113",138
"/display/ATLASSIAN/JIRA+as+CMDB/","37.115.191.27",121
"/display/ATLASSIAN/JIRA+as+CMDB/","46.118.159.224",101
"/display/public/HealthCheckPage","77.243.52.139",5732
"/display/slangereden/","5.9.155.37",118
"/display/slangereden/","66.249.64.19",140
```

To **exclude** these IP Addresses:

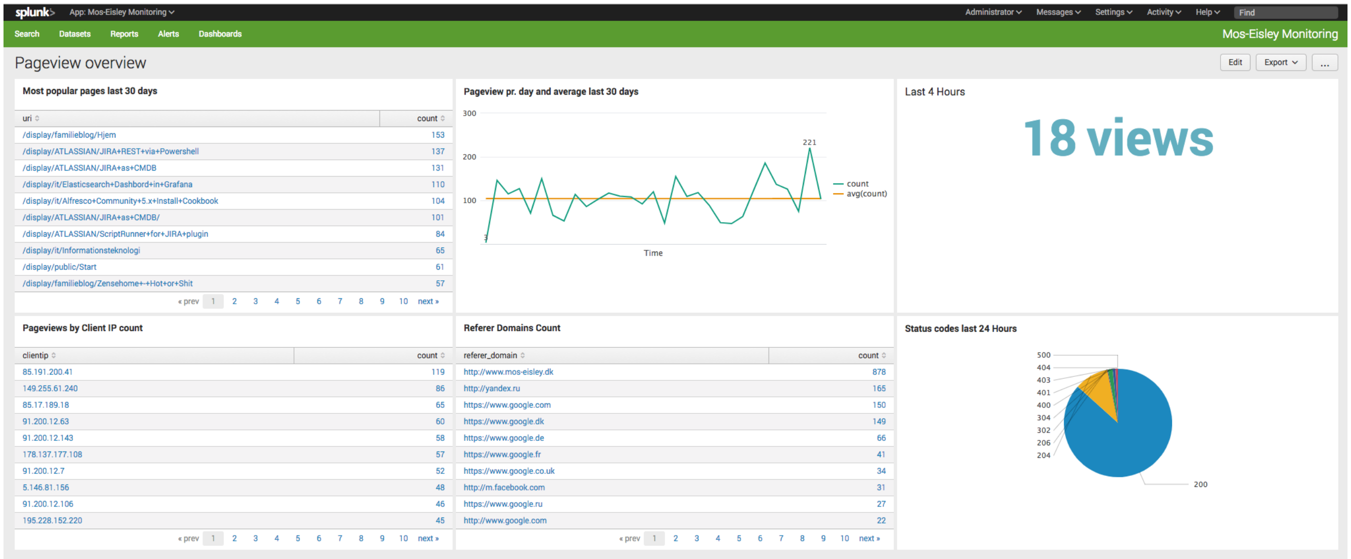
```
NOT [| inputlookup hard_hitting_hosts.csv | fields clientip]
```

To sum up - Conclusion

The final result is this splunk search:

```
(uri="*/display/*" OR uri="*/viewpage.action/*") uri!="/display/public/HealthCheckPage" useragent!="*bot*"
useragent!="*spider*" useragent!="*facebookexternalhit*" useragent!="*crawler*" useragent!="*Datadog Agent*"
NOT [| inputlookup robot_spiders.csv | fields clientip] NOT [| inputlookup hard_hitting_hosts.csv | fields
clientip]
```

Gives a more correct Dashboard:



Pageview overview

Most popular pages last 30 days

uri	count
/display/familieblog/Hjem	153
/display/ATLASSIAN/JIRA+REST+via+Powershell	137
/display/ATLASSIAN/JIRA+as+CMDB	131
/display/it/Elasticsearch+Dashbord+in+Grafana	110
/display/it/Alfresco+Community+5.x+Install+Cookbook	104
/display/ATLASSIAN/JIRA+as+CMDB/	101
/display/ATLASSIAN/ScriptRunner+for+JIRA+plugin	84
/display/it/Informationsteknologi	65
/display/public/Start	61
/display/familieblog/Zensehome+-+Hot+or+Shit	57

« prev 1 2 3 4 5 6 7 8 9 10 next »

Pageview pr. day and average last 30 days

Last 4 Hours

18 views

Pageviews by Client IP count

clientip	count
85.191.200.41	119
149.255.61.240	86
85.17.189.18	65
91.200.12.63	60
91.200.12.143	58
178.137.177.108	57
91.200.12.7	52
5.146.81.156	48
91.200.12.106	46
195.228.152.220	45

« prev 1 2 3 4 5 6 7 8 9 10 next »

Referer Domains Count

referer_domain	count
http://www.mos-eisley.dk	878
http://yandex.ru	165
https://www.google.com	150
https://www.google.dk	149
https://www.google.de	66
https://www.google.fr	41
https://www.google.co.uk	34
http://m.facebook.com	31
https://www.google.ru	27
http://www.google.com	22

« prev 1 2 3 4 5 6 7 8 9 10 next »

Status codes last 24 Hours