

CN REST - Story #8749

Fix log aggregation events from the CN without associated CN IPs

2018-11-16 20:39 - Chris Jones

Status:	New	Start date:	2018-11-16
Priority:	Normal	Due date:	
Assignee:	Dave Vieglais	% Done:	0%
Category:		Estimated time:	0.00 hour
Target version:			
Story Points:			
Description <p>The robots list used to filter out usage events includes the IP addresses of the CNs, so events logged during synchronization don't show up as true hits. Because of the SSL infrastructure at lbl.gov, the ESS-DIVE group doesn't see the public IP of an incoming request, but rather an internal private IP assigned by lbl.gov infrastructure. You can see the impact of this on the ESS-DIVE profile page. The spike of 11,000+ downloads in August 2018 was the CN synchronizing content.</p> <p>Rushiraj summarized these events in a gist</p> <p>There are multiple 10.42.x.x IP associated with the CN requests. These events all need to be updated in the logsolr core and changed to an actual CN IP. For future synchronizations, perhaps we need to add 10.42.0.0/16 to the robots list?</p>			

History

#1 - 2019-02-21 23:03 - Jing Tao

Since those events should be filtered out. So maybe we just delete them? The criteria is the subject is a CN and IP address is 10.42.x.x.

#2 - 2019-02-21 23:12 - Chris Jones

I think it's fine to delete them Jing, since we know they are CN events. They can be deleted from Elastic Search as well, so ask Rushi or Dave about that if need be.

#3 - 2019-02-22 00:12 - Jing Tao

Run this query
curl "http://localhost:8983/solr/event_core/select?q=subject:CN=urn\:node\:CN%20AND%20ipAddress:10.42.*&fl=subject,ipAddress"

It returned 32642 records.

#4 - 2019-02-22 03:02 - Jing Tao

Proposed three three delete command:

```
curl http://localhost:8983/solr/event_core/update/?commit=true -H "Content-Type: text/xml" -d "<delete>(subject:CN=urn\:node\:CNUCSB1*)AND(ipAddress:10.42*)</delete>"
```

```
curl http://localhost:8983/solr/event_core/update/?commit=true -H "Content-Type: text/xml" -d "<delete>(subject:CN=urn\:node\:CNORC1*)AND(ipAddress:10.42*)</delete>"
```

```
curl http://localhost:8983/solr/event_core/update/?commit=true -H "Content-Type: text/xml" -d "<delete>(subject:CN=urn\:node\:CNUNM1*)AND(ipAddress:10.42*)</delete>"
```

#5 - 2019-02-22 18:36 - Jing Tao

This query `curl -d "q=(subject:CN=urn\:node\:CNUCSB1*)AND(ipAddress:10.42*)&fl=subject,ipAddress" http://localhost:8983/solr/event_core/select` returns 32462 records;

`curl -d "q=(subject:CN=urn\:node\:CNORC1*)AND(ipAddress:10.42*)&fl=subject,ipAddress" http://localhost:8983/solr/event_core/select` returns 0 records.

`curl -d "q=(subject:CN=urn\:node\:CNUNM*)AND(ipAddress:10.42*)&fl=subject,ipAddress" http://localhost:8983/solr/event_core/select` returns 180 records.

So the delete queries will totally remove 32,642 records. Chris, does it sounds reasonable number?

#6 - 2019-04-19 17:31 - Chris Jones

Hi Jing - we discussed this with ESS-DIVE yesterday, and it reminded me of this ticket - sorry for the delayed response.

I wanted to get a sense of how many read events your query entailed, so I issued this query:

```
curl -d "q=(subject:CN=urn\:node\:CNUCSB1*)AND(ipAddress:10.42*)&rows=0&facet=true&facet.field=event&facet.limit=1000000" http://localhost:8983/solr/event_core/select | xmlstarlet fo
```

This summarizes the count of each event name, and we get:

```
<int name="updateSystemMetadata">79878</int>
<int name="read">13937</int>
<int name="synchronization_failed">221</int>
<int name="INSERT">0</int>
<int name="UPDATE">0</int>
<int name="create">0</int>
<int name="delete">0</int>
<int name="replicate">0</int>
<int name="unknown">0</int>
<int name="update">0</int>
<int name="upload">0</int>
```

So a large part of the query deletes updateSystemMetadata events and it also catches the synchronization_failed events. I don't think we want to delete those events since they are there for reference, but we also don't want them to have the wrong IP address.

To clean this up, I'd probably say your delete query should be `(subject:CN=urn\:node\:CNUCSB1*)AND(ipAddress:10.42*)AND(event:read)`, and then we probably want to update the remaining Solr documents where `subject:CN=urn\:node\:CNUCSB1*` and change the IP address to the actual IP address, and do the same for the other CN's records as well.

#7 - 2019-05-01 20:06 - Jing Tao

Proposed three delete command:

```
curl http://localhost:8983/solr/event_core/update/?commit=true -H "Content-Type: text/xml" -d
"<delete><query>(subject:CN=urn\::node\::CNUCSB1*)AND(ipAddress:10.42*)AND(event:read)</query></delete>"
```

```
curl http://localhost:8983/solr/event_core/update/?commit=true -H "Content-Type: text/xml" -d
"<delete><query>(subject:CN=urn\::node\::CNORC1*)AND(ipAddress:10.42*)AND(event:read)</query></delete>"
```

```
curl http://localhost:8983/solr/event_core/update/?commit=true -H "Content-Type: text/xml" -d
"<delete><query>(subject:CN=urn\::node\::CNUNM1*)AND(ipAddress:10.42*)AND(event:read)</query></delete>"
```

#8 - 2019-05-01 20:54 - Jing Tao

This page give some information to update a document:

<https://solr.pl/en/2012/07/09/solr-4-0-partial-documents-update/>

#9 - 2019-05-01 22:25 - Jing Tao

Chris and I used the command to delete those records:

```
curl http://localhost:8983/solr/event_core/update/?commit=true -H "Content-Type: text/xml" -d
"<delete><query>(subject:CN=urn\::node\::CN*)AND(ipAddress:10.42*)AND(event:read)</query></delete>"
```