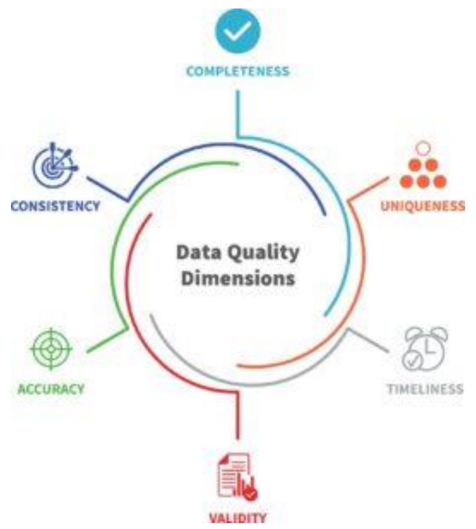
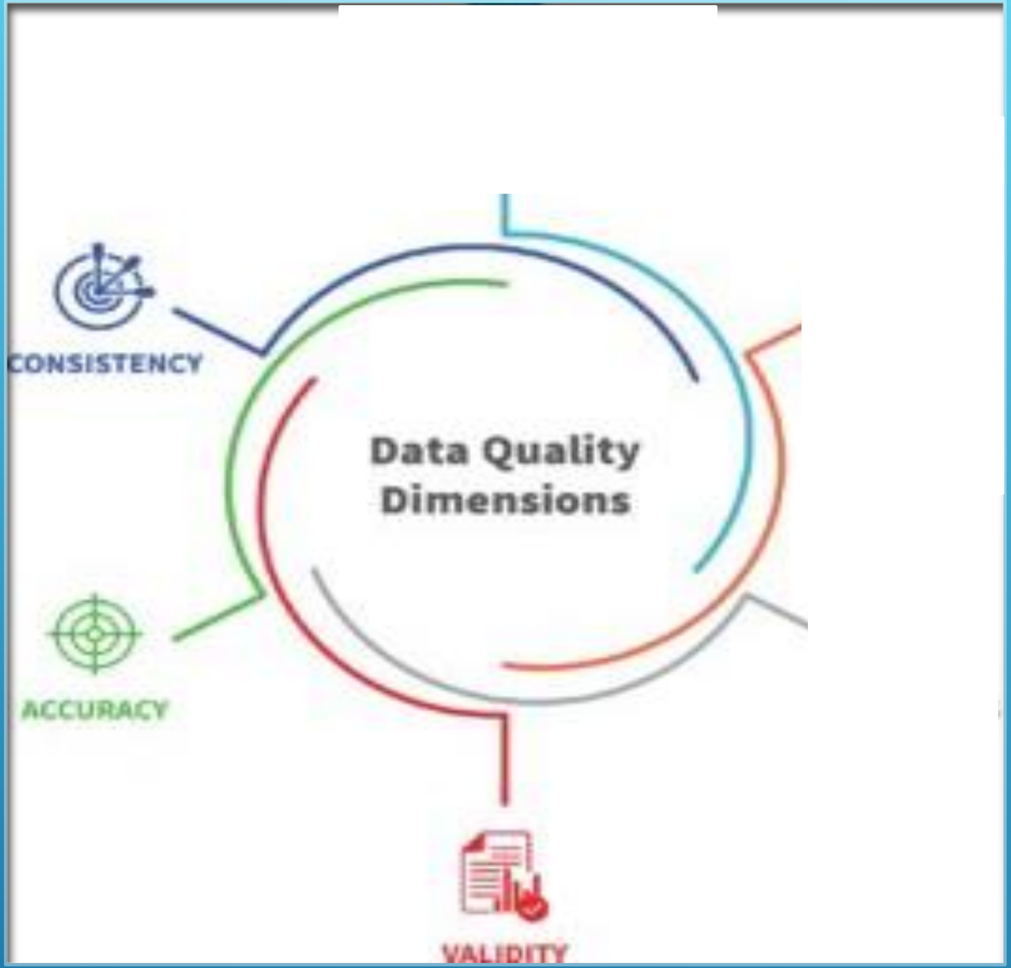
Harrison Dewhurst

DATA QUALITY

PURPOSE



- Data Quality
- Potential future problems
- Need best data quality to perform analytics

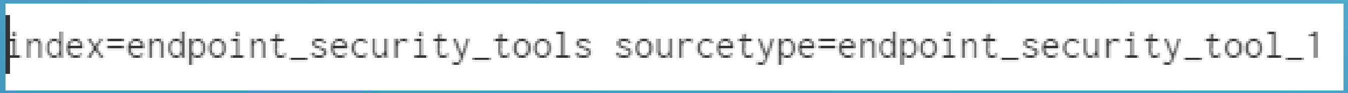

CONSISTENCY
Data Quality Dimensions
ACCURACY
VALIDITY

OVERVIEW

- Eight terabytes of data per day
- Data is kept in a Splunk cloud repository
- Cyber security data from hundreds of sources

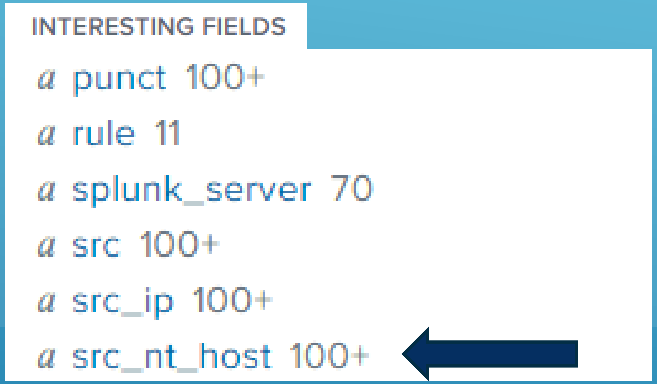
FROM THE BEGINNING

Initial search



```
index=endpoint_security_tools sourcetype=endpoint_security_tool_1
```

Field Name

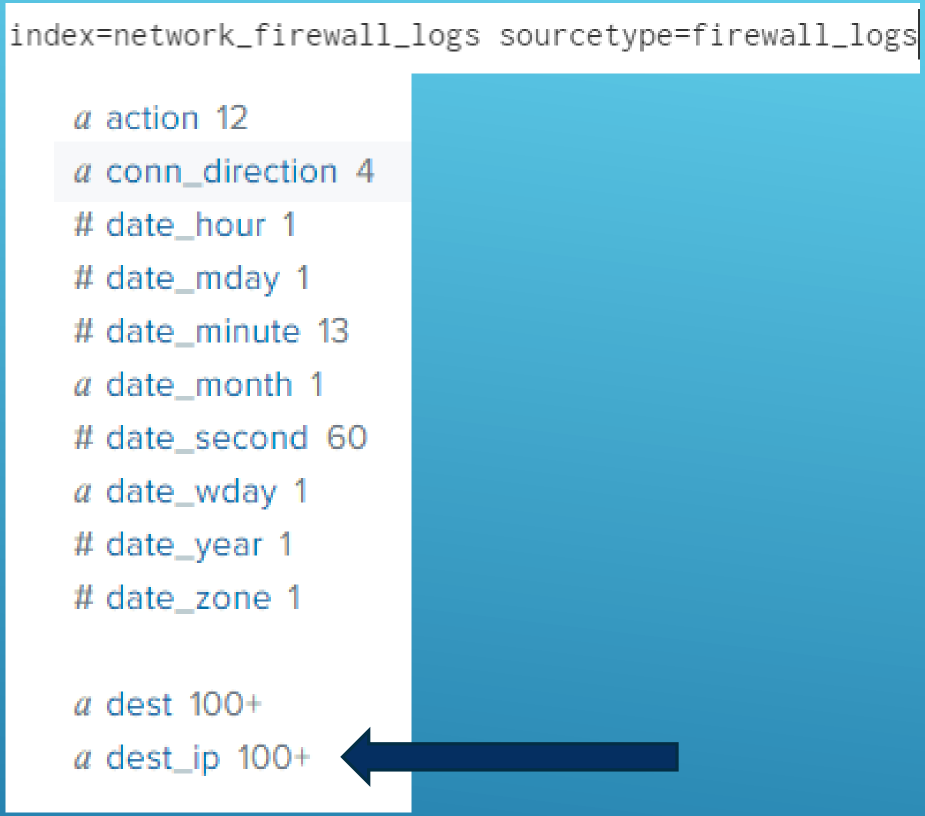


INTERESTING FIELDS

- *a* punct 100+
- *a* rule 11
- *a* splunk_server 70
- *a* src 100+
- *a* src_ip 100+
- *a* src_nt_host 100+

DEST_IP EXAMPLE

Initial search for dest_ip field



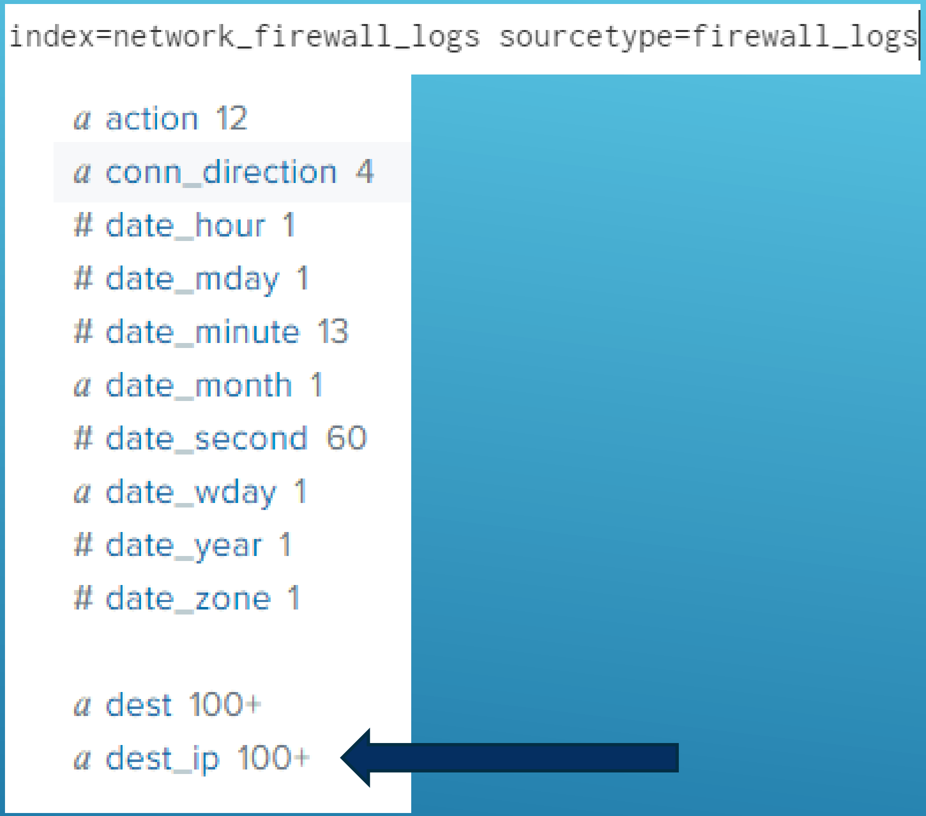
```
index=network_firewall_logs sourcetype=firewall_logs
```

```
a action 12
a conn_direction 4
# date_hour 1
# date_mday 1
# date_minute 13
a date_month 1
# date_second 60
a date_wday 1
# date_year 1
# date_zone 1

a dest 100+
a dest_ip 100+
```

DEST_IP EXAMPLE

Initial search for dest_ip field

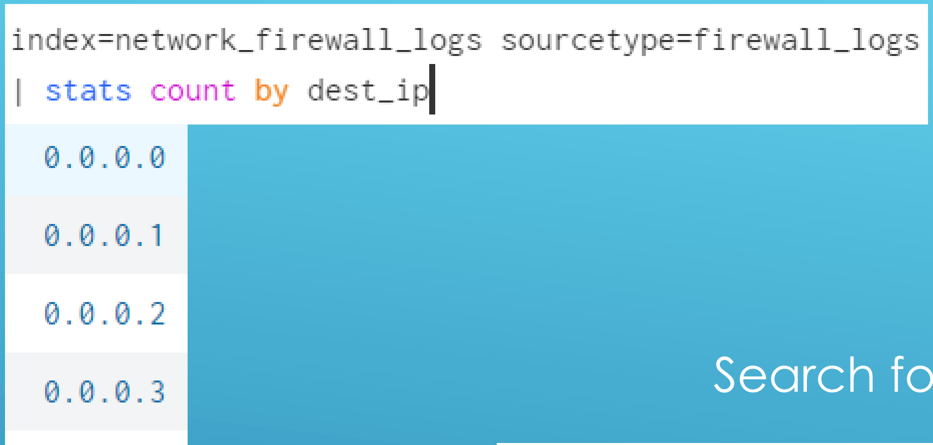


```
index=network_firewall_logs sourcetype=firewall_logs
```

```
a action 12
a conn_direction 4
# date_hour 1
# date_mday 1
# date_minute 13
a date_month 1
# date_second 60
a date_wday 1
# date_year 1
# date_zone 1

a dest 100+
a dest_ip 100+
```

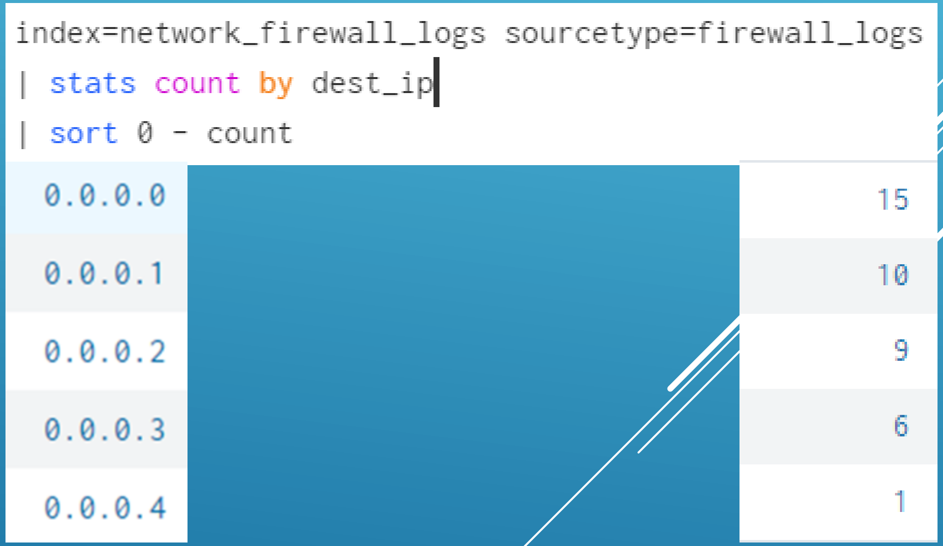
Look at fields for initial impurities



```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by dest_ip
```

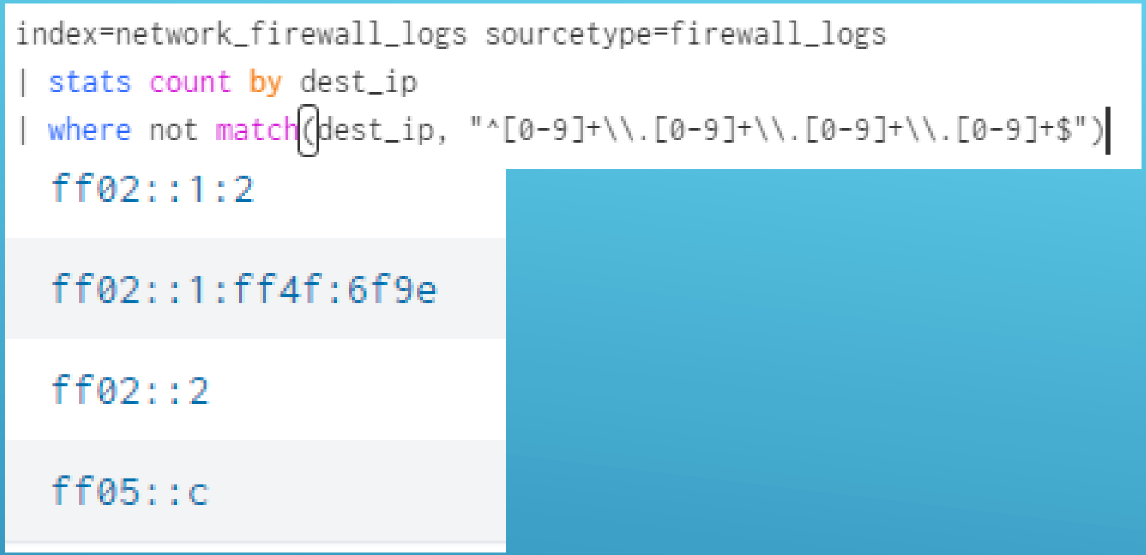
```
0.0.0.0
0.0.0.1
0.0.0.2
0.0.0.3
0.0.0.4
```

Search for defaults



```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by dest_ip
| sort 0 - count
```

0.0.0.0	15
0.0.0.1	10
0.0.0.2	9
0.0.0.3	6
0.0.0.4	1



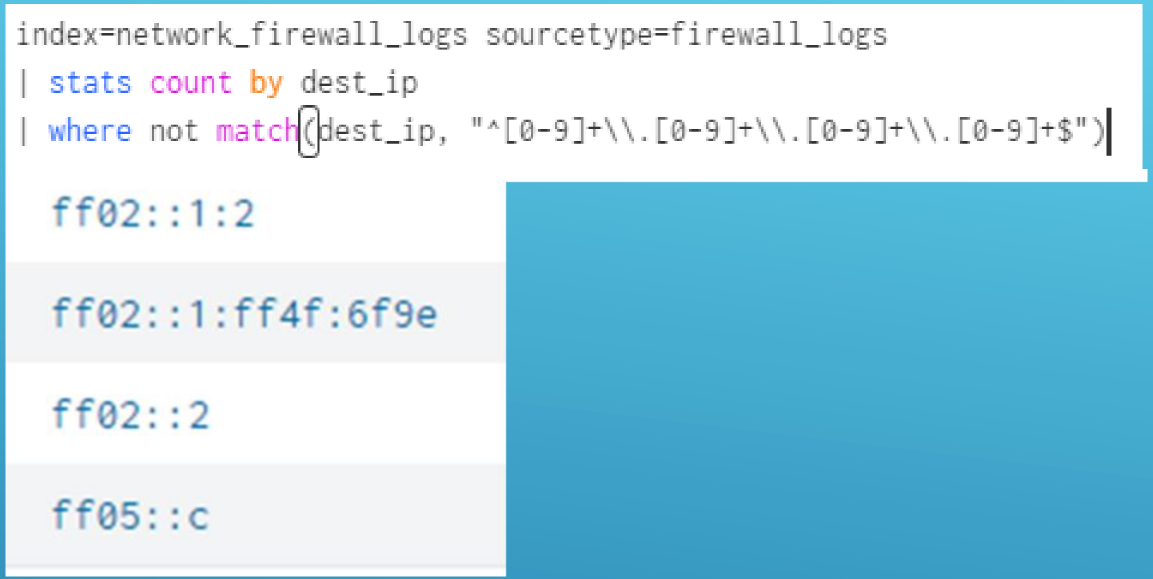
```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by dest_ip
| where not match(dest_ip, "^[0-9]+\\.[0-9]+\\.[0-9]+\\.[0-9]+$")
```

```
ff02::1:2
ff02::1:ff4f:6f9e
ff02::2
ff05::c
```

ADVANCED INVESTIGATION

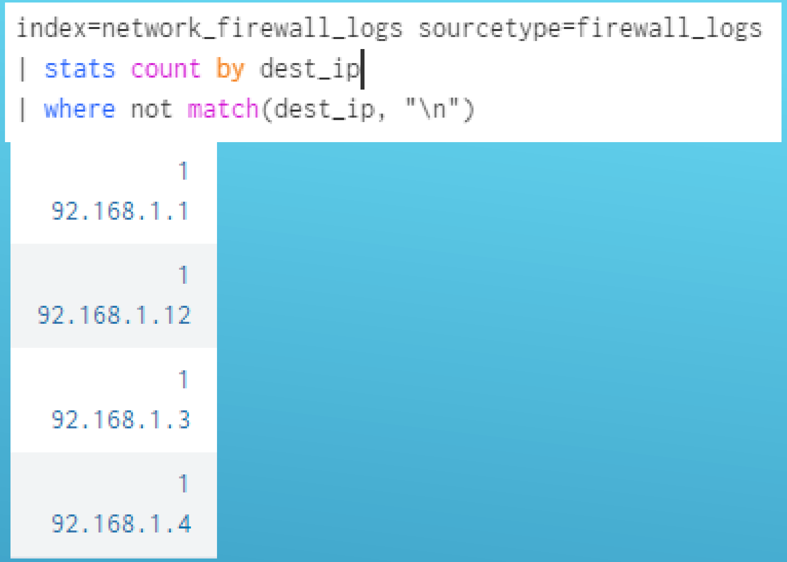
ADVANCED INVESTIGATION

Regex query for non IPV4 events



```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by dest_ip
| where not match(dest_ip, "^[0-9]+\\.[0-9]+\\.[0-9]+\\.[0-9]+$")
```

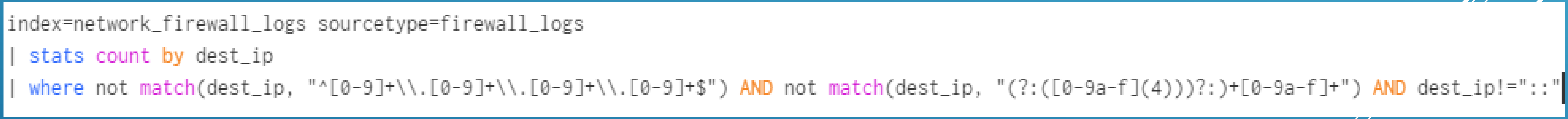
```
ff02::1:2
ff02::1:ff4f:6f9e
ff02::2
ff05::c
```



```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by dest_ip
| where not match(dest_ip, "\n")
```

```
          1
 92.168.1.1
          1
92.168.1.12
          1
 92.168.1.3
          1
 92.168.1.4
```

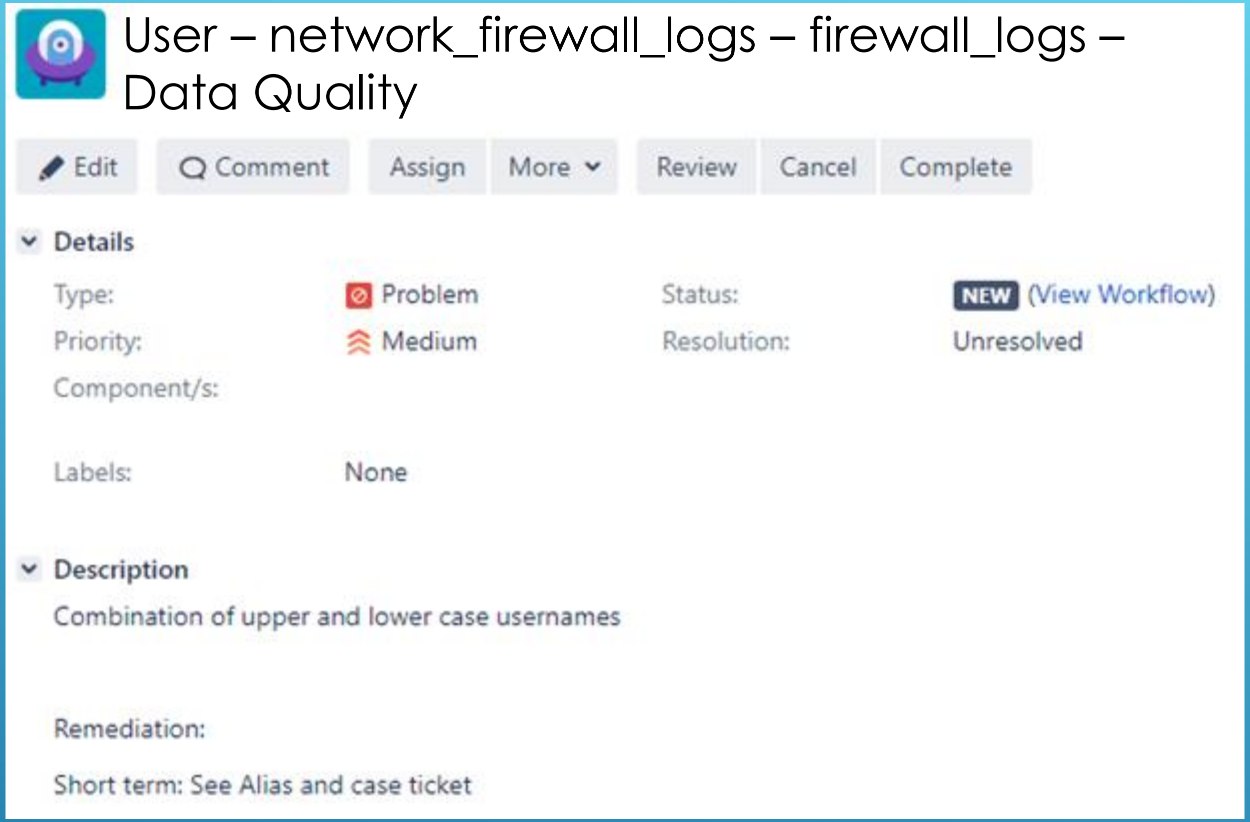
Advanced search for non IPV4, non IPV6, and no double colons



```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by dest_ip
| where not match(dest_ip, "^[0-9]+\\.[0-9]+\\.[0-9]+\\.[0-9]+$") AND not match(dest_ip, "(?:([0-9a-f](4)))?:)+[0-9a-f]+") AND dest_ip!="::"
```

DETAILED DOCUMENTATION

Data Source	Firewall
Index	Network_firewall_source
Sourcetype	Firewall_1
Has dest_ip field	Yes
Destination IP Address Field	dest_ip
Contains Non Numeric Characters?	No
Follows IPV4 Standard Format?	Mostly
Defaults	No
Standard	No
Comments	Some events are almost IPv4 events but at some point in the ip address there is a return and the rest of the ip address is continued on the next line.



User – network_firewall_logs – firewall_logs – Data Quality

Details

Type:	Problem	Status:	NEW (View Workflow)
Priority:	Medium	Resolution:	Unresolved
Component/s:			
Labels:	None		

Description

Combination of upper and lower case usernames

Remediation:

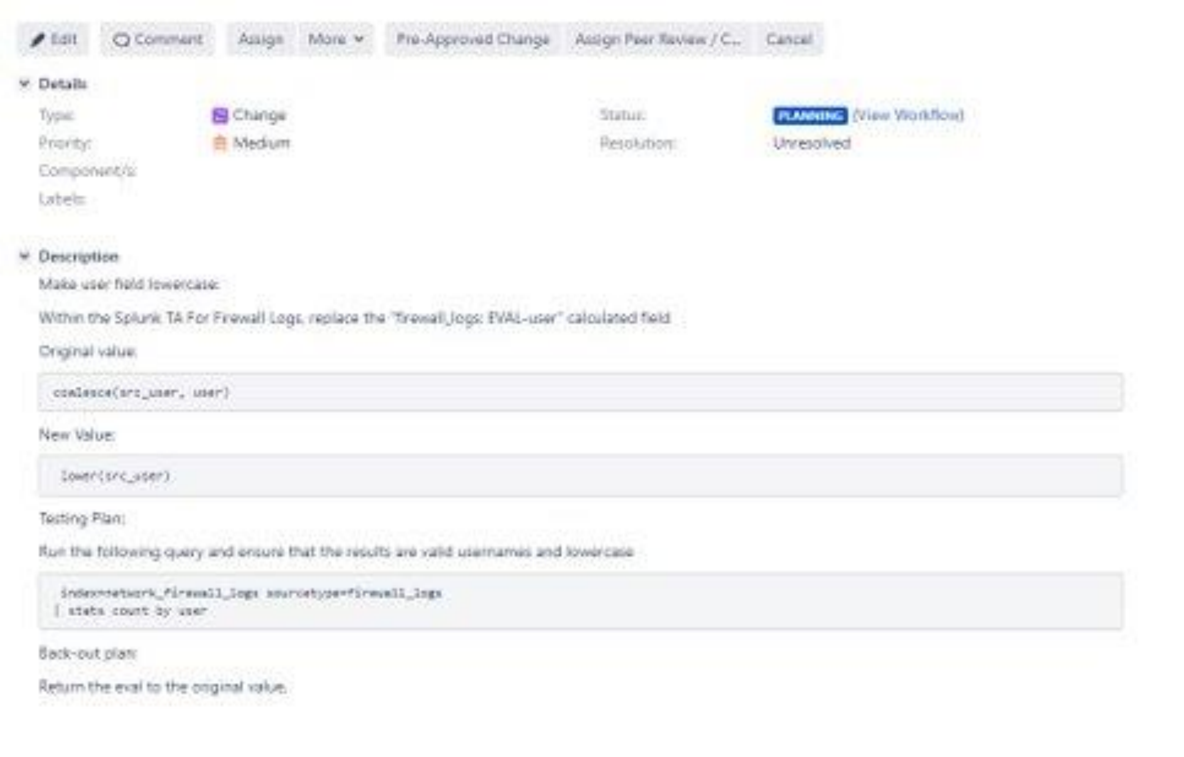
Short term: See Alias and case ticket

PROBLEM TICKETS

- Description of problem and comments from investigation performed

CHANGE TICKETS

User – network_firewall_logs – firewall_logs – Data Quality



Details

Type: Change

Priority: Medium

Component/s:

Labels:

Status: PLANNING (View Workflow)

Resolution: Unresolved

Description

Make user field lowercase:

Within the Splunk TA For Firewall Logs, replace the "firewall_logs: EVAL-user" calculated field

Original value:

```
coalesce(src_user, user)
```

New Value:

```
lower(src_user)
```

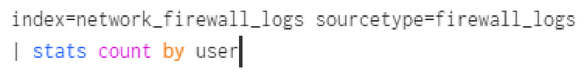
Testing Plan:

Run the following query and ensure that the results are valid usernames and lowercase

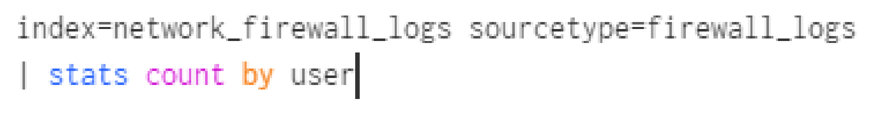
```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by user
```

Back-out plan:

Return the eval to the original value.



```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by user
```



```
index=network_firewall_logs sourcetype=firewall_logs
| stats count by user
```

Results Redacted

- Implement change in Splunk

QUESTIONS?

Thanks for listening!