IETF Hackathon

Personal Information Identification in Logs

IETF 106

16-17, Nov 2019

Singapore

Hackathon Plan

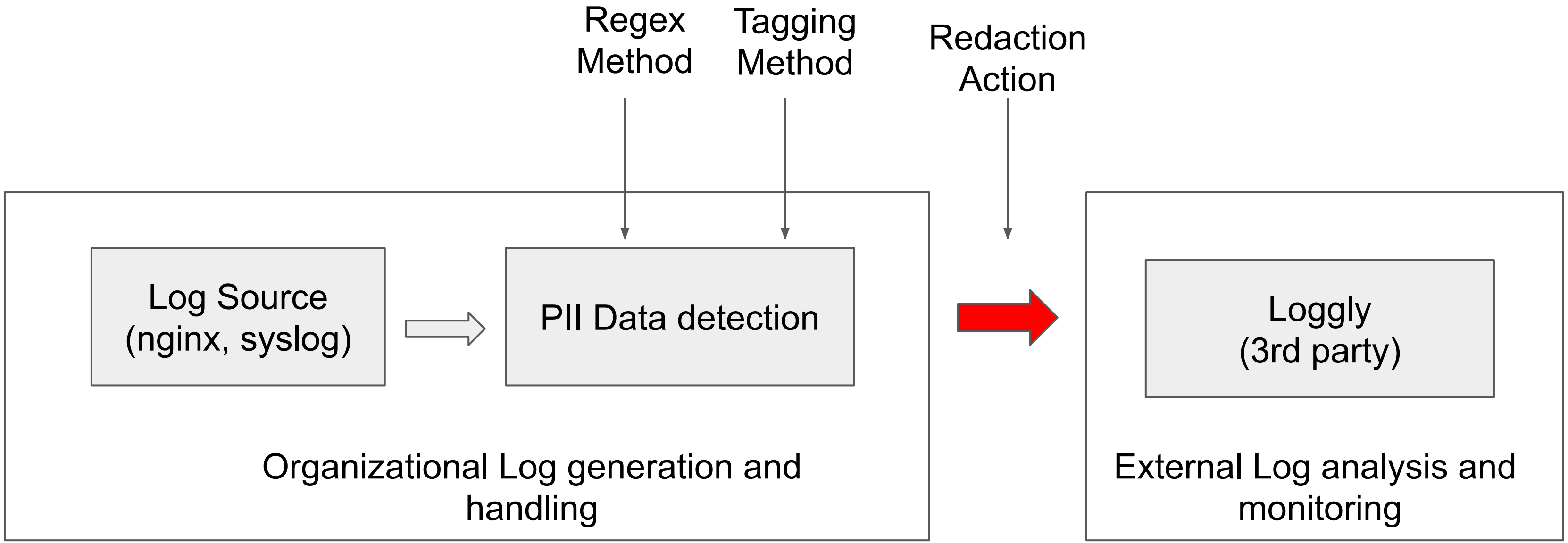
Objective

- Address need for personal data identification in system logs

Scope

- NGINX and Syslog formats - scalar and structured log data
- Compare methods of regex / dictionary with explicit personal data tagging
- Display of redacted data on Loggly

Log Pipeline



Regex Method

Tagging Method

Redaction Action

Log Source (nginx, syslog)

PII Data detection

Organizational Log generation and handling

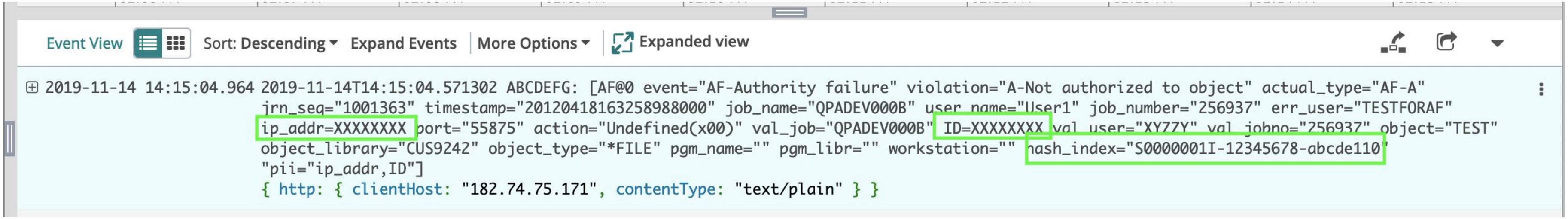
Loggly (3rd party)

External Log analysis and monitoring

Log Input

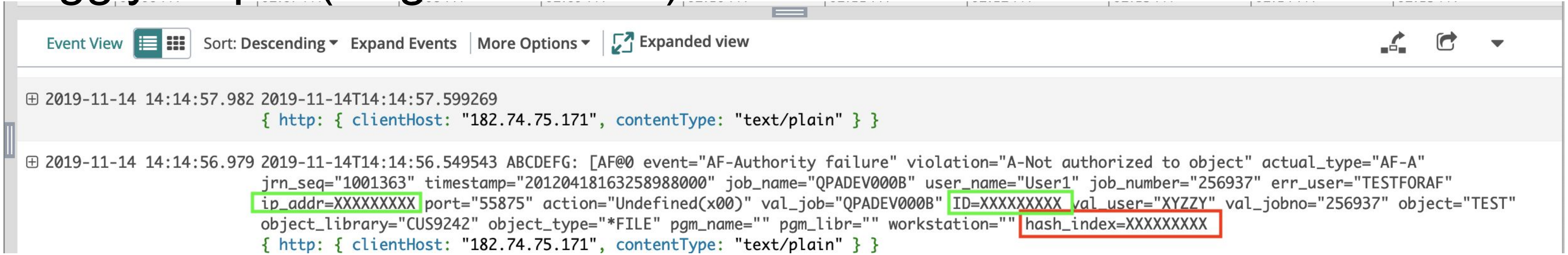
```
ABCDEFG: [AF@0 event="AF-Authority failure" violation="A-Not authorized to object" actual_type="AF-A" jrn_seq="1001363"
timestamp="20120418163258988000" job_name="QPADEV000B" user_name="User1" job_number="256937" err_user="TESTFORAF"
ip_addr="10.0.1.21" port="55875" action="Undefined(x00)" val_job="QPADEV000B" ID="S0000001I" val_user="XYZZY"
val_jobno="256937" object="TEST" object_library="CUS9242" object_type="*FILE" pgm_name="" pgm_libr="" workstation=""
hash_index="S0000001I-12345678-abcde110" "pii="ip_addr,ID"]
```

Loggly Ouput (Tagging Method)



```
⊞ 2019-11-14 14:15:04.964 2019-11-14T14:15:04.571302 ABCDEFG: [AF@0 event="AF-Authority failure" violation="A-Not authorized to object" actual_type="AF-A"
                          jrn_seq="1001363" timestamp="20120418163258988000" job_name="QPADEV000B" user_name="User1" job_number="256937" err_user="TESTFORAF"
                          ip_addr=XXXXXXXX port="55875" action="Undefined(x00)" val_job="QPADEV000B" ID=XXXXXXXX val_user="XYZZY" val_jobno="256937" object="TEST"
                          object_library="CUS9242" object_type="*FILE" pgm_name="" pgm_libr="" workstation="" hash_index="S0000001I-12345678-abcde110"
                          "pii="ip_addr,ID"]
                          { http: { clientHost: "182.74.75.171", contentType: "text/plain" } }
```

Loggly Ouput (Regex Method)



```
⊞ 2019-11-14 14:14:57.982 2019-11-14T14:14:57.599269
                          { http: { clientHost: "182.74.75.171", contentType: "text/plain" } }

⊞ 2019-11-14 14:14:56.979 2019-11-14T14:14:56.549543 ABCDEFG: [AF@0 event="AF-Authority failure" violation="A-Not authorized to object" actual_type="AF-A"
                          jrn_seq="1001363" timestamp="20120418163258988000" job_name="QPADEV000B" user_name="User1" job_number="256937" err_user="TESTFORAF"
                          ip_addr=XXXXXXXXX port="55875" action="Undefined(x00)" val_job="QPADEV000B" ID=XXXXXXXXX val_user="XYZZY" val_jobno="256937" object="TEST"
                          object_library="CUS9242" object_type="*FILE" pgm_name="" pgm_libr="" workstation="" hash_index=XXXXXXXXX
                          { http: { clientHost: "182.74.75.171", contentType: "text/plain" } }
```

What we learnt

- Regex based detection has potential for false positive and misclassification
- Explicit Personal Data tagging at source is more effective
- Challenges
  - Structured -vs- Unstructured log format
  - Scalar log data vs AV log data