# Project Overview

Creation of a smart tool to automate or semi-automate the data extraction and normalization process at Telstra.

## Project background

In Telstra, large and varied data sets are used to detect security threats and anomalies. These data sets pertain to all types of logs ranging from access logs to network & security logs.

The data, generated from multiple technologies comes in different shapes as the result of varying

- Naming conventions
- Configurations
- Logging standards
- File formats

> For example, Telstra's next generation firewall data has a number of vendors in this space (Palo Alto, Cisco, Fortinet) each one of them have different log formats with similar data.

## Normalization is the pre-requisite to all analysis.

Logs that contain unknown format or are a currently unmapped format can not be analysed by current 3rd party integrations (Splunk, ElasticSearch etc).

The data in these logs needs to be normalised into a valid format after which it can be fed into existing 3rd party integrations allowing for their security analysis.

**Current Solution**

The current solution requires human effort to understand the logs. This is followed by manually mapping fields from previously unknown formats to a common schema.

### This can be characterised as:

- Time consuming
- Prone to human errors.

### Project Motivation

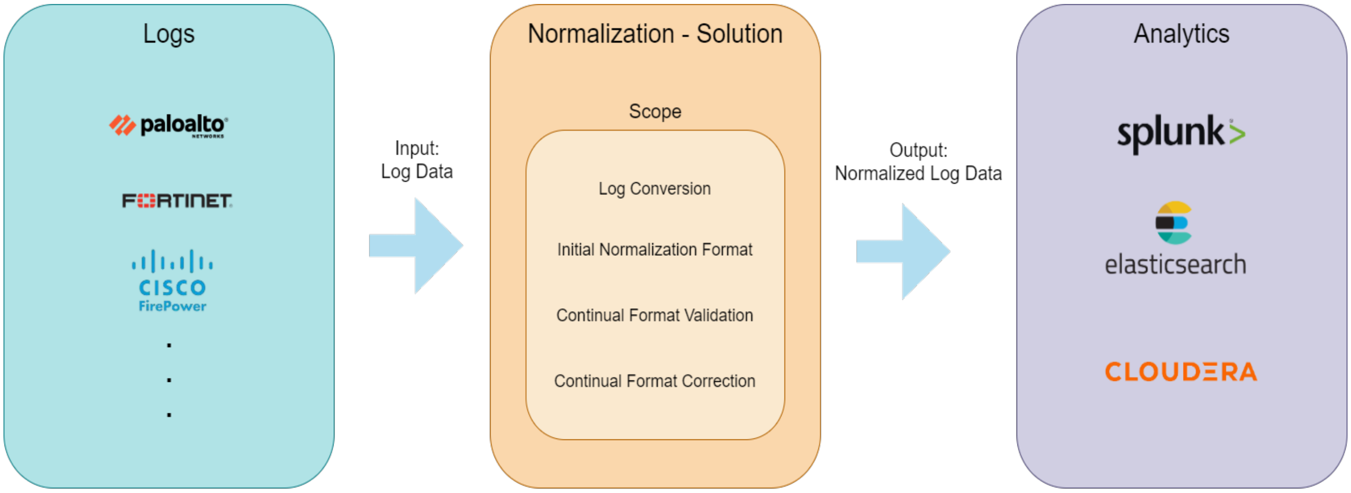
The motivation for this project is to reduce the time expenditure on log normalisation, improve error margins and simplify the cluttered nature of the current process.

### This project will produce value to the business and can be characterised as

1. Reduction of human hours spent on manual normalisation.
2. Ingestion of more ever-changing data for security analysis.
3. Increase ability to secure Telstra's Systems, as this normalised data can now be analysed by their 3rd party integrations (Splunk, ElasticSearch)
4. Mitigating the costs of the security breaches from the enormous amount of confidential data Telstra handles.

## Project goal

To engineer a consistent approach to data normalization.



Logs

Input:
Log Data

Normalization - Solution

Scope

Log Conversion

Initial Normalization Format

Continual Format Validation

Continual Format Correction

Output:
Normalized Log Data

Analytics

## Not in Scope:

- Log Data Ingestion
- Analytics

## Example of analytic integrations used at Telstra

Elastic ECS

Splunk CIM

## Examples of Logs used at Telstra

Palo Alto (most recent format)

Fortinet

Cisco Firepower

# Project Details

## Core issue

Telstra currently takes in terabytes of log files from firewalls, load balancers, servers, and other network devices on a daily basis. Once collected, these log files are parsed and are then ingested by analytics platforms to be analysed in real time. As log specs vary heavily between device manufacturers /providers, data fields and contents must be normalised to allow for meaningful analysis and comparison to take place. Currently, this normalisation process requires significant manual effort – we are tasked with investigating possible routes towards creating an automated or semi-automated normalisation.

## Current Solution:

### Primary users:

Telstra data engineers and security analysts that would have in the past had to investigate and determine mappings themselves.

### Secondary users:

Data scientists / analysts / existing automated ML platforms that ingest the processed data.

### Current process:

1. Take a ~1000 line file
2. Engineer will manually map data fields based on either:
    a. Logging specification or interface agreement provided by logging vendor (not always available)
    b. Guesswork
3. Engineer creates parser to convert data into a normalized format for consumption/analysis.
4. Output normalized logging file is a well formed data table, works well for a lot of end user.

The software and systems used to clean and normalise the log data are listed below:

- Custom built parsers/mappers that process logs from 90+ device types.
- ELK stack (Elastic w/ Logstash) with custom configs for normalisation.
- Spunk + SplunkES with custom technical addons alongside some prebuilt common information model mappings that have been provided by vendors.

### Pain points:

- Problem only occurs infrequently, only when adding new software. So building a solution might not be worth it.
- Logging specs aren't always available or may be out of date, and some new logging features added by vendors may not be adequately documented.
- Different methods of data extraction
- Determining mappings is labour intensive and can be quite complex. It has been described as "as much art as science".
- Introduce error from manual human work
- Vendors do not always notify Telstra when updates change the format or content of log files. Similarly, others teams may not always notfiy the the integration and tuning team around when they plan to install updates.

### Output consumers:

| Technology | Use Volume | Priority |
|---|---|---|
| Splunk | 1000's | High |
| Elastic Search | 100's | High |
| Cloudera | Unknown | Medium |

- Splunk uses the Splunk Common Information Model, while Elastic uses the Elastic Common Schema
- 3 different teams, inconsistency in output standards

### Input consumers:

| Vendors | Format | Format Notes | Example |
|---|---|---|---|
| Palo Alto | value, value, value, value, value... | Values separated by commas, field depends on position | FUTURE_USE, Receive Time, Serial Number, Type... |
| Fortinet | Field=value Field=value Field =value... | Field specified with value by equals sign and separated by space | date=2017-11-15 time=11:44:16 logid="0000000013" type="traffic"... |

| | | | |
|---|---|---|---|
| Cisco Firepower | Field = value, Field = value.... | Field specified with value by equals with spaces and separated by comma | Group = groupname, Username = user, IP = IP_address ... |
| Many more vendors | | | |

Note: Same vendors can have different logging formats based on configuration

### Input scale and format:

| Formats | Volume | Devices |
|---|---|---|
| Many, JSON, XML, CSV, Txt | 3-5Tb a day | 10,000's, Routers, Servers, Endpoints |

## Solution Expectations:

- Literature review / survey of current open source or commercially available tools that could solve problem.
    - Outlining tools, value, pros, cons, cost, and configuration settings.
    - Common practices in industry
    - Analysis of effectiveness. For example, discussion and comparison around the mapping accuracy of:
        - Rule-based matching
        - Machine learning or statistical inference models
        - Existing mapping processes (may require that Telstra provides data)
- A design for a tool with an interactive interface that can output recommended mappings / formats that an engineer can then review and approve or override
- The solution should be able to predict mappings for unknown fields and recognise standard formats for hostnames, usernames, and other common fields.
- It is desirable if the solution is "smart" and learns from existing mappings.

The client has expressed no strong preference regarding how the tool is delivered - they are open to web solutions, desktop applications etc.

### Scope:

- Our solution should provide recommendations towards partially automating the normalization of log data from a variety of log formats and vendors.
- Our solution should be capable of producing output to be analysed by Splunk or Elasticsearch.

### Out of Scope:

- No system ingest requirement, just normalization.
- The solution should not be fully automated - processes for manual review must be included as part of the system. Client confirmed that they would not trust a fully automated solution.

### Areas of Extension:

- Our proposal could include methods for automatically identifying changes to log and data formats provided by existing systems/network devices and adapting normalization processes to reflect these changes.
- Our proposal could include features for assessing the effectiveness of a proposed mapping by testing it against a large body of data. For example, the solution could provide coverage data or other statistics for a given mapping or regex.