Project Overview

About the Company

As Australia's biggest telecommunications provider, Telstra handles a large majority of internet, mobile, and landline connections in the country.

In recent years there has been a big shift from offline to online services, especially with the introduction of the 5g network and NBN.

Because of the increase in users on their networks, Telstra has a responsibility to ensure that the networks are secure to protect their users.

To address this issue, all of Telstra's infrastructure records data logs, which contain security information that is analysed to assess threats to their networks.

Problem

Telstra's hardware infrastructure records data logs, which needs to be interpreted and standardized into a format before being analysed on 3 primary data platforms: Splunk, Cloudera, and Elastic, which Telstra's security teams can use to analyse threats and anomalies.

The infrastructure hardware varies greatly, from servers, routers, laptops etc., in addition to the different operating systems running on the hardware.

With this comes two problems:

- data being outputted in different file formats (e.g xml vs json)
- data features being named differently, with some formats including more/less features in the log file (e.g userId vs user_id)

These factors make converting the log files for input into a database, and subsequent analysis, a difficult problem. There have been previous efforts in the past to standardize messages from the message generating software but have fallen short as it requires all manufacturers to conform to the standard. Rather than attempting to deal with the logs at the application output source, it is more feasible to introduce a normalization process to extract common fields and convert them into a standard format.

The product work is commissioned by Telstra staff (Dave Wilson, Terence Chen) and will be used by multiple data teams within Telstra.

The key stakeholders in this project are:

- Telstra Leaders (Dave Wilson, Terence Chen)
- Telstra Staff (data teams, security teams)

System

Current Solutions to the Problem

Fields and columns are being mapped manually before take into the data analysis systems(Splunk, Elastic, Cloudera). This process is labour intensive and prone to human error.

Existing Systems

There are several alternative tools for normalizing log data, such as normalization with rsyslog. However, they are not a comprehensive solution, which contains several limitations:

- Not automated or semi-automated.
- Normalization speed is too slow.
- Not sufficient for security data because it isn't real-time; security logs need to be analysed in real-time.
- Lack of automatic detection of format changes from the data source.

Desired Solution

It is important to note the client has intentionally given vague specifications for the proposed solution. This has allowed for an open ended interpretation by the requirements team to decide the type of system to implement.

The final solution will consist of two parts:

1. An assistive tool that can assist with the normalisation of data to a more consumable form. It will provide suggestions regarding what a data field is, and allow for data exploration.
2. Documentation of normalisation processes. Involves a review of industry standard data normalisation practices.

The benefits of the system will be:

- **Telstra Data Teams** workflows involving the data will be streamlined by using the tool; minimisation of complexity, frustration, and other negativities of the current process.
- Security threats and anomalies will be caught more accurately as there is less room for human error, benefiting **Telstra Security Teams**.

Scope of the solution

In Scope

This project intends to provide a solution to assist with normalisation of varied data formats. Several tasks were identified as part of the proposed scope of a solution to meet the basic requirements of the design problem.

- The tool to be built for input to a system and has an interactive interface to provide recommended data fields, and formats for data views, so that all level of users can easily access and work with the tool conveniently.
- The tool will observe data with different formats and naming conventions, consider standard forms for common data fields, and inform the user of patterns.
- Detection of unexpected format changes from the source is a beneficial addition. With sufficient abstraction, it would allow the tool to be adaptable to changes with source system file formats.
- The tool should have the capacity to deal with 3-5 TB of data daily, and different file formats and types of log files: sys log, database log, firewall log and event log.

Out of Scope

- High-level methods like inference and machine learning models can be adopted in future enhancements, using correct mapping percentages as a performance metric.
- The tool is not expected to automatically or semi-automatically create a format of data by mapping source data to a more consumable format, such as a database with columns and rows. An automated system would not be trusted.
- Preference is not to change the ingestion process because Splunk and Elastic have their own.
- Data routing. Receiving and outputting data will not be handled by the program, only the normalisation is relevant.

Reference

- Elastic: https://www.elastic.co/guide/en/ecs/current/ecs-field-reference.html
- Splunk: https://docs.splunk.com/Documentation/CIM/4.18.0/User/Overview
- Rsyslog: https://www.rsyslog.com/log-normalization-for-different-formats/
- https://www.researchgate.net/publication/310545144_Efficient_Normalization_of_IT_Log_Messages_under_Realtime_Conditions#pfc