Data Sources

Log Files

For our prototype we used two sample log files, one for structured data and one for unstructured data.

The unstructured data is from Loghub's sample of Linux logs, specifically lines 339-356. These lines were chosen as they have a number of IP addresses and dates in the data, some simple pieces of structured data that can be easily captured.

The structured data comes form the Australian Goverment's dataset website and is a list of the Colac Otway Shire Trees. To reduce the width of the data sample, the shorter lines of lines 2-26 were selected.

When a regular expression is built and saved, it will be stored in a relational database as such:

ID	Regular_Expression	Sample_Log_Entries	Type	Vendor	Created_By_ID	Created_By_Name	Date_Created	Last_Modified
1	(?<linux_ftpd_server_date>\w+ \d{2} \d{2}:\d{2}:\d{2}) .* (?<linux_server_code>\w+\[\d{5}\]).* (?<ip_address>\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})	Jun 25 09:20:24 combo ftpd [31475]: connection from 210.118.170.95 () at Sat Jun 25 09:20:24 2005 Jun 25 09:20:24 combo ftpd [31477]: connection from 210.118.170.95 () at Sat Jun 25 09:20:24 2005 Jun 25 09:20:24 combo ftpd [31474]: connection from 210.118.170.95 () at Sat Jun 25 09:20:24 2005 Jun 25 09:20:24 combo ftpd [31476]: connection from 210.118.170.95 () at Sat Jun 25 09:20:24 2005	Traffic	Cisco	353578	Terence Chen	02/04/2021	28/04/2021