2021/03/16 - Client Meeting

Attendees

- Dave - Client
- Terrance - Client
- Renata - Supervisor
- Team Koala
- Mustafa Awni
- Soth Bak
- Ben Nguyen
- Tina Tang
- Michael Thomas
- Sayyaf Waseem

Agenda

- Permission to record
- Introductions - Team and Client
- Introduction of System
- Team Questions
- Client Questions and Expectations
- Any further information & wrap up

Questions

Initial

- What problem are we looking to address?
    - Telstra has a large security center that keeps Telstra secure and also sells services to customers.
    - Telstra has log file data coming in, stored for analysis. Data can only be analysed after it has been normalised, currently a manual process, and is a bottle neck.
    - Every device in the network (laptop, endpoint tools on user laptops, router, switch, server, firewall, software stack, load balancer, AWS, Cloudwatch, database logs, authentication systems, windows, event logs) sends log files into "several very large repositories". Each device has their own format, and needs to be adjusted to be compatible
        - Over 100,000 network devices sending unstructured and unnormalised data
    - Current system has many input files, and once processed it goes to 3 primary platforms: Splunk (and Splunk ES), Elastic, and Cloudera
- What are the objectives of the system to be built? What do you expect the system to accomplish?
    - Document industry practices, and create a tool to automate the process
    - Leverage pre-existing industry standards, algorithms, and solutions.
    - The output of our product should be processible by Spunk, Elastic and the cloud application

Understanding the System As Is

- Could you describe the system as is? How is this problem being solved today?
    - Starts in data ingestion, where the source data is gathered. Data is then passed into Splunk & Elastic where the data is processed.
    - Currently, data fields are being labelled manually
    - Way to figure out how to process is to find the device's specification, which lays out what the device records
    - In Spunk & Elastic the data is processed, usually via a lookup, and educated guess.
        - Elastic is used for a product that Telstra sells
        - Splunk is used internally (at a larger scale)
- What are the pain points?
    - Data is received in different file formats such as CSV, JSON, XML.
    - Sometimes the format of a log file from a device changes, so ideally the system should be able to detect these changes
    - Data must be normalised & ingested in order to be manipulated. Currently this is a manual effort.
        - When the log files go to all 3 systems (Splunk, Elastic, Cloudera), each team (3 total) will manually name columns and fields in their own way
            - Labour intensive and prone to human error
    - Servers are running a new OS that hasn't been seen before. It has log files that are relevant from a security perspective. User behaviour data is recorded e.g. logged in, logged out, did abc. Data is helpful for security to see if something odd is happening. Ingesting the log file is worthless though because they don't know what a user ID looks like.
        - Need to extract common fields that have security value given the files are from different vendors with different formats. Need for normalisation.
- Is the entire process manual?
    - At the moment, the creation of the way of processing a new input file is manual
- Who are the users of this platform?
    - Primary Users: Data Analysists & Data Scientists, with varying level of expertise, from entry-level (such as graduate) to expert
    - Secondary Users: People who use the results of the system. Data is fully normalised, ingested, formatted, outputted when it reaches these users.
        - Also cybersecurity team has 10-20 users who may access final data
- Could you show us an example of the system's Input and Output?
    - Not at the moment, at a later interview
- What level of scale should the end product be able to handle?

    - Number of peak users - just a 'handful', as the other analysists use the results, not the system itself
    - Amount of daily data ingest - about 3-5 TB of data daily
    - Types of log files - sys log, database log, firewall log, event log
- What is the priority of the output services, relative to each other?
    - Spunk/Elastic, then Cloudera last (need to contact Patrina from Telstra to find the top priority)
    - For each output service, there is a team dedicated to it, with their own processes.

Understanding the System To Be

- What is the preferred application type? (Web, Mobile, Desktop, Existing Tech)
    - Unsure at the moment, something for us to think what is best
- What business environment will this system be used in?
    - Analysist/Engineer will parse data, validate recommendations, then pass data into the product, and get the formatted data out
- What is the nature of system to-be?
    - An automated or semi-automated tool
    - Tool can be inputted to system and has interactive interface to provide recommended format for data views after normalisation
    - At this point of time, system to-be is just to create format of data. Preference is not to change ingestion process (Splunk and Elastic have their own).
        - Maps out source data to a more consumable format like a database with column and rows. Objective is to line up values up into columns. (Currently all done manually)
    - Detection of unexpected format changes from the source.
        - Primary concern is to map out format for normalisation, but beneficial to periodically check if anything has changed with source system file formats
    - Platform/system agnostic
- What are the initial thoughts on what may be wanted from SWEN90009?
    - A paper prototype
    - Proposal or recommendation. Different approaches to address problem.
    - Interface agreement for the three Telstra teams using data. Example:
        - Step 1 - A script which separates out values. Parse data into consumable format.
        - Step 2 - Standardise the fields (different vendors call them differently).
            - This would require manual lookup from documentation, or guessing the meaning of a data item
    - A review or literature survey of tools that are already solving similar problems
        - What is the value of these tools: Pros and Cons
        - Analysis on how effective the tools are. May need to use Telstra's data for this analysis
        - Inference and ML model, seeing how much mapping we are doing as percentage successfully

Notes

- No recording of start, as host wasn't in attendance at start
- Future meetings - A good idea, as some other members of Telstra's team has more information
- Of the two clients: Terrence is lead technical person rather than Dave Wilson
- Client wanted to know what we'd be delivering - Requirements, User Stories, Paper Prototype, End of Sprint Meetings
- Terrence doesn't have any leave planned for the future, so he should be available.
- Currently aiming for weekly meetings with client, with both teams coordinating to prevent asking the client something twice, but can change to fortnightly if it's too many

Action Items

- [ ] Turn information into requirements
- [ ] Book next meeting