1. Introduction

2. Splunk

Splunk is a software tool for searching, monitoring and analysing machine generated data via web interface. It indexes and correlates real-time and non-real-time big data to generate meaningful statistics and visualizations. It is used for application management, security and compliance as well as business analytics.

2.1 Input

Splunk reads textual data through various methods to perform its indexing. Splunk can read and index any form of textual data. Splunk reads data and divides the data into events based on the timestamps on the data. If the real time data does not contain timestamp, Splunk puts the timestamp when the data is read or puts a timestamp of the data saved in the file. Splunk however cannot read binary data. In case of binary data, user needs to input a way to convert the binary data to textual information. This can be done by running scripts along with splunk to convert the binary data.

2.1.1 Types of input

- Files or directories

  Files or directories can be given as an input for Splunk indexing. Different files can be given under the same data source so that splunk monitors the files all the time and indexes real time based on the updated file.
- Reading from UDP / TCP Ports

  Splunk can read data from UDP / TCP connections. Splunk can listen to the specified ports and data can be read real-time and indexed.
- Running custom scripts

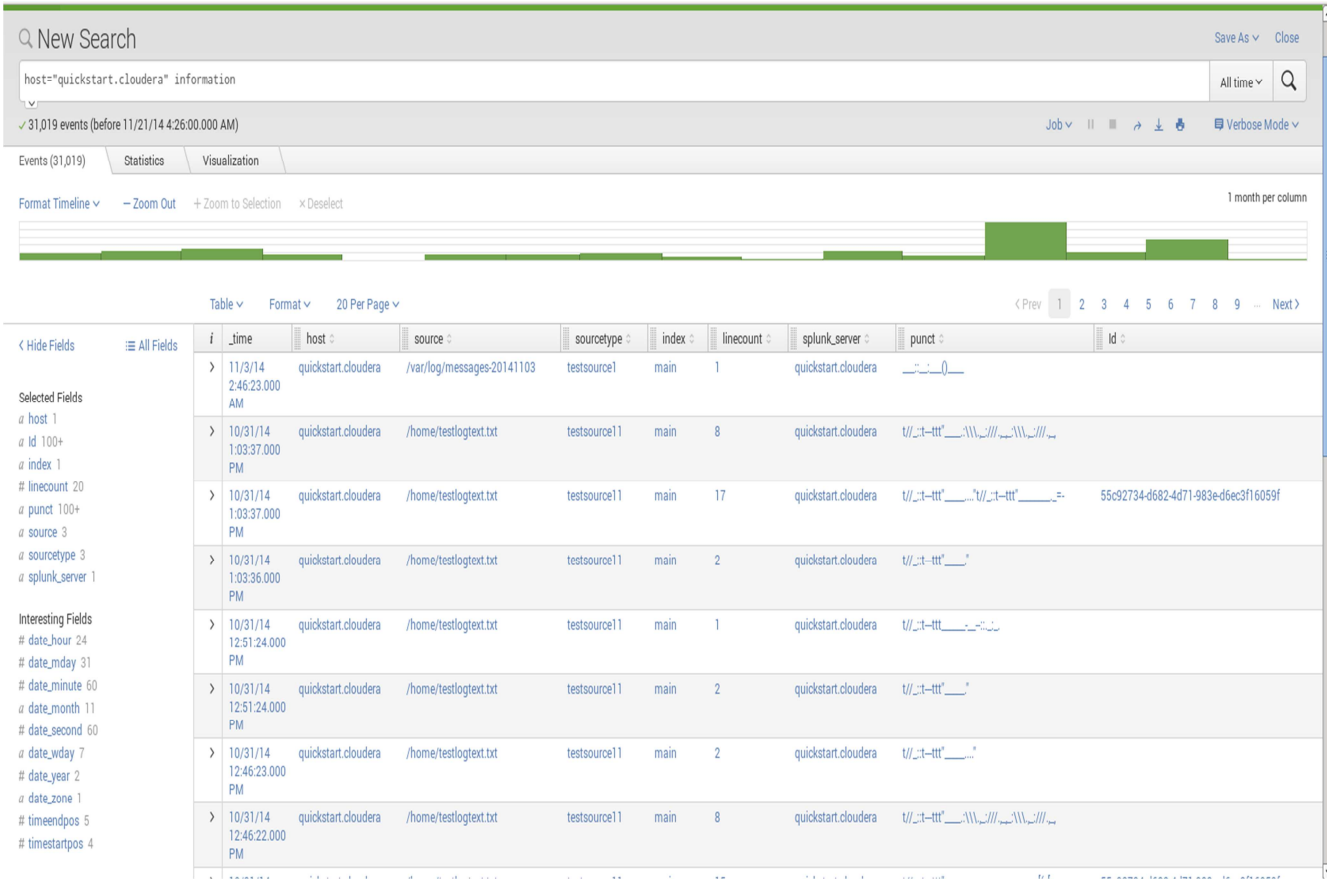
  Splunk can read data from script outputs or program outputs.

2.2 How splunk indexes data

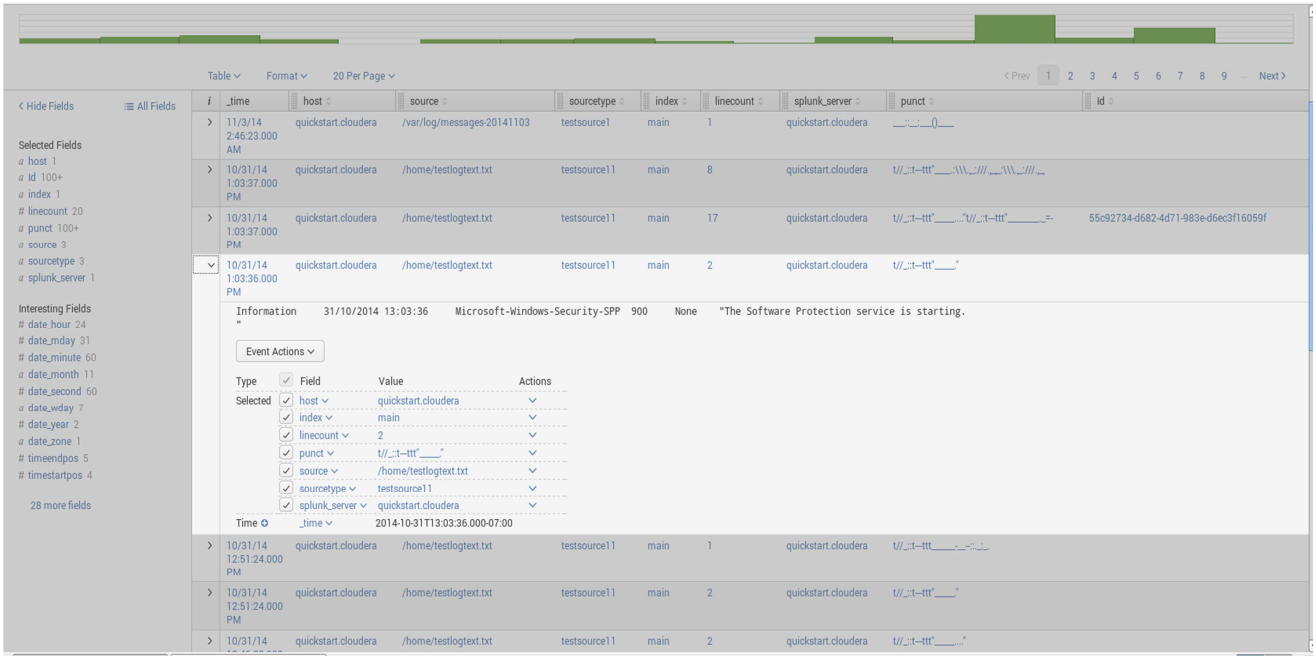
Splunk reads machine data from its corresponding input method and indexes the raw data based on the timestamp. It creates a time based mapping of the data. The data is divided into individual events each possibly separated by a timestamp. An individual can vary from a single line log to millions of lines. The search criteria are run on these individual events.

For each event, splunk maps the data to the inbuilt fields in splunk. The inbuilt fields cover major information like timestamp, source, source type. Additional fields can also be added to the inbuilt fields by modifying the configuration files.

Li Zhao, IT Innovation Specialist
itinnovation@contacts.bham.ac.uk



Indexing



2.2.1 Searching using splunk

Splunk uses a search process language to find, filter, summarize and visualize the data. The search process language is an SQL like query language.

Setting up search criteria is essential in terms of analysing security threats and log analysis. The data can be filtered from millions of files depending upon the source, type of source, keywords, timestamps or user added fields. The visualization and statistics views of splunk give a better idea about the frequency of the threats or errors in the log files. This helps in building mechanisms to fight the threats and errors from the applications.

2.2.2 Dashboards

Dashboards are search panels that can be saved and used later for future use. Dashboards works in the same way as search in splunk. Dashboard has additional features in the form of controls. Controls include Radio buttons, checkboxes and dropdowns. These controls help to modify the search criteria depending upon their values making the searching easier than using search queries. These are search criteria which can be saved and re-used at any point of time. These search criteria can be edited as well which gives it enough flexibility to add or remove options for search.

2.3 Recipes and Monitoring

Splunk allows creating recipes and monitoring certain users based on the log data. The application can add certain conditions called recipes to identify ineffective users and monitor them. Alerts can also be scheduled based on this monitoring system.

3. Splunk app development

Splunk has an inbuilt web framework to develop apps for specific purposes. Splunk apps consist of dashboards and UI control panels catered for the user's requirements. The dashboard editor is built in simple xml which makes it easier for developers to use it efficiently as well as create light weight apps. It also gives options of developing using JavaScript and Django. Alternatively, Splunk also has SDKs for popular programming languages to start app development from scratch.

Splunk apps can also be developed to be add-ons to other applications. Add-ons can include custom search commands, field extraction, source type definitions etc.

4. Splunk apps

Splunk has various apps development by the company as well as individual users for security threat analysis and log analysis. Listing some of the apps useful for the context

- **Splunk app for enterprise security**
  The splunk app for enterprise security identifies and addresses the emerging security threats by monitoring and analysing the information. The features include incident review, automated correlated searches, reports and security metrics, risk based analysis, threat intelligence framework. This is a paid app.

- **Logfiller app**
  The log filler application is a free application developed to discover slow websites, applications, analyses VDI performances and security issues.

- **Splunk app for Unix and Linux**
  The splunk app for UNIX and Linux is a free application developed to identify the performance and capacity bottlenecks in the UNIX and Linux environment.

- **Tripwire enterprise app**
  Tripwire enterprise app visualizes the health of the IT environment with rich data, controls and policies. This reduces the cycle-time of identifying vulnerabilities of the applications. This is a paid app.

5. Experiment

Setup

Splunk recommends setting up the software for evaluation purposes in Cloudera or Hortonworks which are standard Virtual Machines with Hadoop instance. This experiment uses Cloudera 5.2 with an inbuilt Hadoop instance for running the splunk.

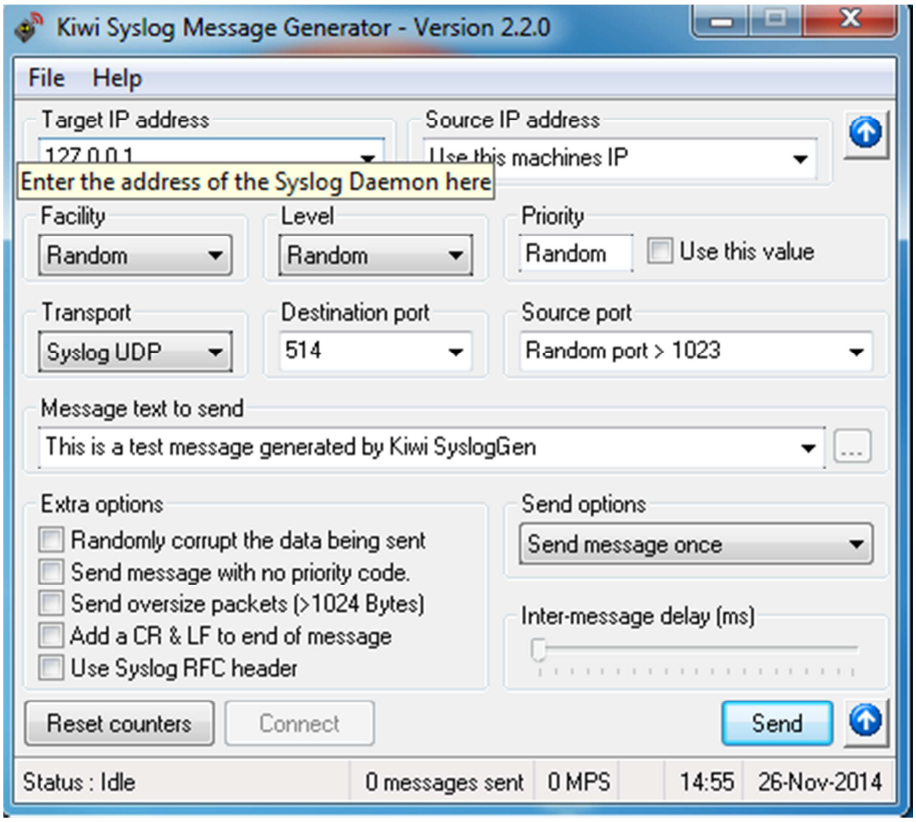
Experiment with Splunk sample data

Splunk provides sample data which is manually loaded to the HDFS nodes and a virtual index is setup to be familiar with splunk indexing features.

Experiment with log files and windows event logs

A software called Kiwi syslog generator was used to create random logs and send it to the virtual machine. Kiwi syslog generator sends random syslog to the destination IP address using TCP or UDP connections. The syslog level can be changed and random corrupt data can also be sent using the software. The splunk on the VM listens to the port to which the syslog data is sent to and indexes through the data. Another source of data includes windows event logs. Event logs from the windows system was exported into the VM and a source was created in splunk for the log files.

Splunk sources were setup to read data from the log files as well as to the UDP port so that the corresponding logs can be merged together.



Kiwi Syslog Message Generator

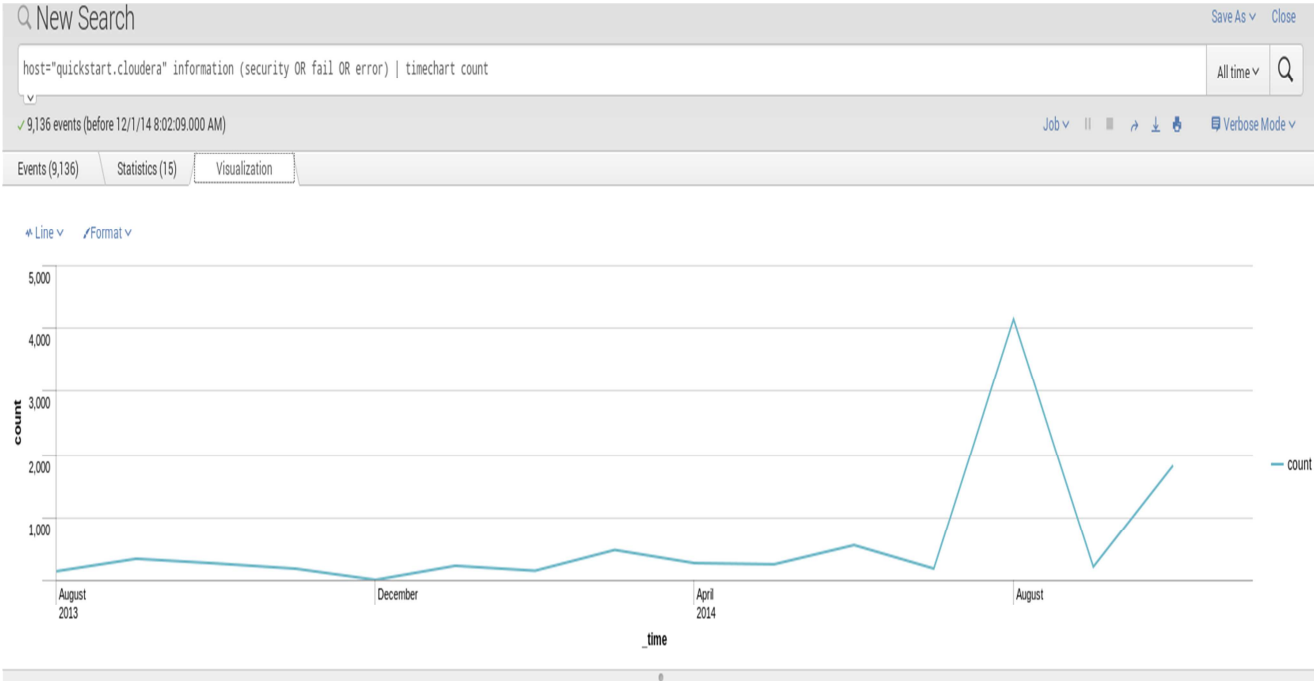
Experiment included the following:

- Indexing search criteria/ Statistics View / Visualization

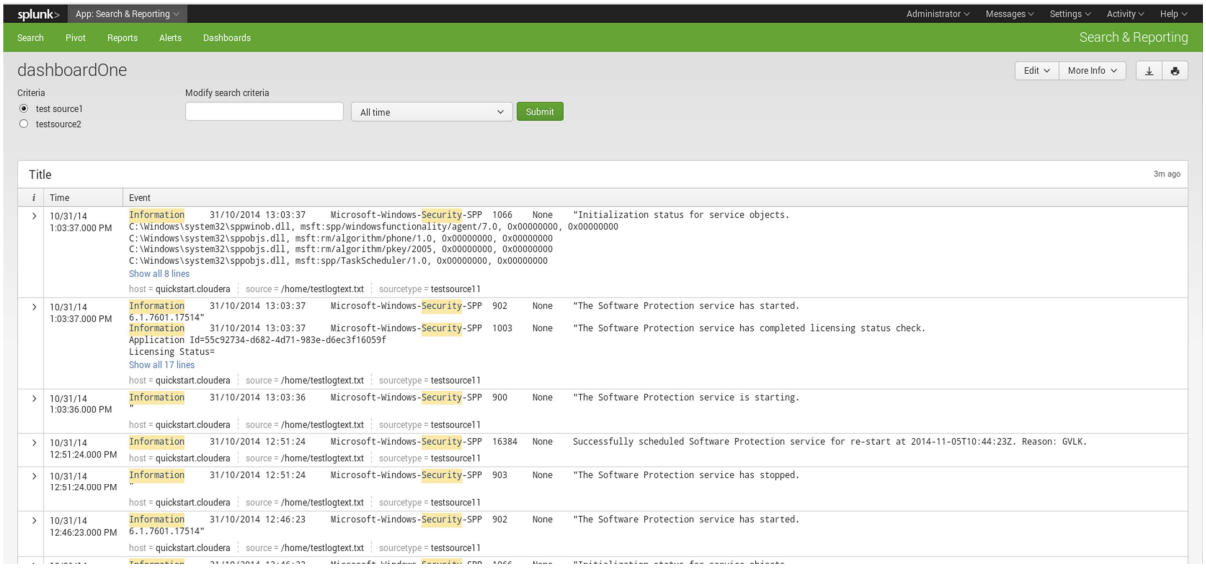
  The events were indexed based on the content of the log message. The information with contents like security or error was filtered and a statistics of the indexed data considering the count of the log information over a certain period of time was measured. This data was also visualized giving a better view of the statistics over that selected period of time.

- Creating dashboards

  Dashboards are searches that can be saved and modified according to certain inputs. A basic search can be given and it can be modified depending upon various input panels like radio boxes, text box, drop downs etc.



Visualization



Dashboard

6. Conclusion and Future Work

Splunk is a great tool for log aggregation and indexing real time server logs and has been widely used in the industry. With several user interactive features and options to tweak the designs according to user's needs, it makes a great product to scan through the logs. The experiments conducted on splunk have been limited based on static logs, windows event logs and syslogs. The server logs would also give a better idea on setting up the search criteria and testing the tool on better experimental conditions. The future work of splunk evaluation lies on running it on real time server log data tagging suspicious IPs, reporting errors instantaneously, monitoring loads and setting alerts for important events. It would also be great to try out various splunk apps or modify certain open source splunk apps according to the needs of the user.

Ref: Splunk Review, Dec 2014
Mohammad Rameez Shafsad
itinnovation@contacts.bham.ac.uk

7. Appendix

Installing Splunk

- Set up the virtual machine and identify its IP address.
- Transfer hunk sample data from hosting system to virtual machine HDFS user home directory. If hosting system is a Windows system, use WinScp else use SCP from command line.
- Follow hunk tutorial from the website. http://docs.splunk.com/Documentation/Hunk/6.1/Hunktutorial/WelcometotheHunktutorial

References & Useful Resources

- http://www.splunk.com/view/SP-CAAAHSM
- https://es.splunk.com/web_assets/v5/book/Exploring_Splunk.pdf
- https://apps.splunk.com/
- http://dev.splunk.com/
- http://www.peppm.org/Products/splunk/price.pdf
- http://repository.cmu.edu/cgi/viewcontent.cgi?article=1713&context=sei&sei-redir=1&referer=http%3A%2F%2Fscholar.google.co.uk%2Fscholar%3Fstart%3D10%26q%3Dsplunk%26hl%3Den%26as_sdt%3D0%2C5#search=%22splunk%22

Mohammad Rameez Shafsad
itinnovation@contacts.bham.ac.uk