**IST-719**

**Week-3 Homework**

Topic: Visually describe a Dataset

Taught by Prof. Gary Krudys

Student: Sharat Sripada (vssripad@syr.edu)

# **TABLE OF CONTENTS**

[**TABLE OF CONTENTS** 2](#_Toc94976142)

[1.0 Introduction 2](#_Toc94976143)

[2.0 About the data 3](#_Toc94976144)

[2.1 Calculating the score of the data 3](#_Toc94976145)

[2.2 Visualizing the data 4](#_Toc94976146)

[2.2.1 Alerts by severity−type in network 4](#_Toc94976147)

[2.2.2 Alerts by node−type in network 4](#_Toc94976148)

[2.2.3 Alert spread across time-of-day 5](#_Toc94976149)

[2.2.4 Dissecting alerts generated by network functions 6](#_Toc94976150)

[2.2.5 Visualizing health of the network 6](#_Toc94976151)

[3.0 Conclusion 7](#_Toc94976152)

# 1.0 Introduction

This homework and all following deliverables for this course will consume a dataset extracted from a network log aggregator, that delivers insights about problems that may exist in on-premises datacenters to network and application administrators. Several nodes and applications generate logs that help developers and administrators root-cause issues and sometimes it becomes important to aggregate them in one place so insights can be drawn across stacks.

In this specific case, logs from the syslog daemon on several nodes like network/security-based control-plane, data-plane (Hypervisors), edges and virtual-machines hosting services like DHCP, DNS etc. have been re-directed to a network log-insight tool. The idea is to use the logs in their raw form and extract interesting insights like the following:

* Alert categories and their distribution by counts *(see Fig1. Plot: Alerts by severity−type in network)*
* Alerts grouped by network nodes that tend to provide an insight into what nodes are generating higher volumes *(see Fig2. Plot: Alerts by node−type in network)*
* Alerts plotted by type across time to understand if there is correlation of higher volumes at certain times of the day *(see Fig3. Plot: Alerts across time-of-day)*
* Grouping network nodes by function and visualizing a distribution across alert types *(see Fig4. Plot: Dissecting alerts generated by network functions)*
* Gathering an overall health of the network at-a-glance *(see Fig5. Plot: Health of the network)*
* A spatial map of the nodes in the network based on hostnames and components (*TBD*)
* Heat map of alerts to indicate hot zones that require immediate attention (*TBD*)

# 2.0 About the data

The data was downloaded in a csv format (separated by ‘,’) and loaded into a data-frame using the read.csv function. A sample line is shown below:

240,"<179> 2021-06-23T19:16:02.291784+00:00 nsx-edge-1-c0 NSX 32192 LB [nsx@6876 comp=""nsx-edge"" subcomp=""nsx-edge-lb.lb"" level=""ERROR""] ""no virtual server defined in lbs ff70393e-4490-471b-8007-8e22b647fa8f.""",2ed1b9c6-7c72-4ced-9078-03401a93eaba,545,82883403,NSX,,nsx-edge,,v4\_dcb8d3e0,local6,/var/log/li-syslog,nsx-edge-1-c0,ERROR,,,LB,22.0,3.0,32192,,,err,20.20.216.157,nsx-edge-lb.lb,,2021-06-23 19:16:02.291,,,,,,,,,,,,,,,,nsx-edge,,,nsx-edge-lb.lb,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,edg lb comp edg subcomp edg lb

Following are some insights of the data:

Graphical user interface, text, application

Description automatically generated

Fig. Code: R code showing structure of data

Graphical user interface, application

Description automatically generated

Fig. Code: R code showing column-names of data

## 2.1 Calculating the score of the data

Using the dim function, the row and column count was obtained as below:

> dim(data)

[1] 323972 78

Now, using the calculator (NumberOfColumns \* 4) \* (NumberOfRows/100) we obtain the following score:

(78 \* 4) \* (323972/100) >> 100

## 2.2 Visualizing the data

### 2.2.1 Alerts grouped by severity in network

Chart, bar chart

Description automatically generated

Fig1. Plot: Alerts by severity−type in network

### 2.2.2 Alerts grouped by node−type in network

Chart

Description automatically generated

Fig2. Plot: Alerts by node−type in network

### 2.2.3 Alert spread across time-of-day

Chart, box and whisker chart

Description automatically generated with medium confidence

Fig3. Plot: Alerts across time-of-day

### 2.2.4 Dissecting alerts generated by network functions

Chart, bar chart

Description automatically generated

Fig4. Plot: Dissecting alerts generated by network functions

### 2.2.5 Visualizing health of the network

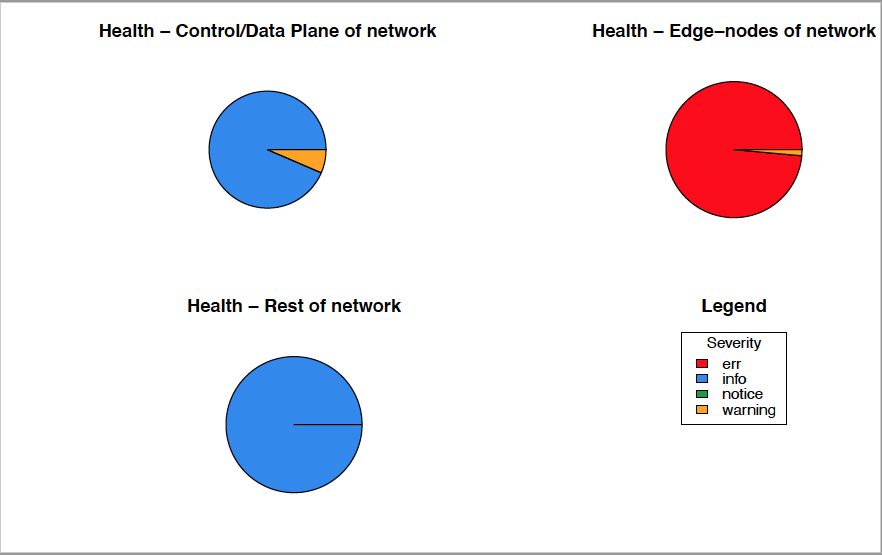


Fig5. Plot: Health of the network

**NOTE:**

A separate PDF comprising all merged plots will be separately attached to the submission.

## 3.0 Conclusion

Using the visualization techniques taught in the first 3-weeks of class IST-719, an unstructured dataset of logs was analyzed to recognize various alert severities and indicate where the network was running hot. Extracting these visual insights from raw data can quickly help dive into the root of the problem and improve critical indices like Mean Time to Recover (MTTR). This is especially useful as we drill from chart to another using APIs or URLs at the back end.

Via visualization plots (especially Fig4 and Fig5), it seems apparent that the Edge-nodes are generating alerts of severity ‘Error’. This would have likely had an impact on network productivity.

In the weeks to come and lining up to the *Final Poster* the scope will expand to identify components within the nodes that caused ‘Error’ alerts. Also, spatial graphs interconnecting network nodes and heat-maps should further alleviate the data story.