Assignment 3: Hadoop, MapReduce, R, and Mahout

Justin Sallese

April 22, 2017

Hive code

Below are data results that follow the analysis that is provided within chapter 3 of the book Information Security Analytics: Finding Security Insights, Patterns, and Anomalies in Big Data by Mark Ryan M. Talabis, Robert McPherson, I. Miyamoto and Jason L. Martin.

Situation: The data that has been used is made up of apache combined server log files. The analysis that is to be done is an analysis of the types of access to the server that the log files have come from. The analysis is to be done useing the hive program that is running on cloudera quick start VM.

Below is the results to the hive query whether or not if there was an attack using SQl injection. Below can be seen the result in which there was an attack useing SQL injection where the attacker tried to access user names and passwords from the database. If this where a proffesional report to a company I, based on the data that is provided would reccomend furthur defenses to be set up so attacks like the SQL injection would not happen again in the future.

SQL Injection Attack

Below are the result to the hive query which was checking if there where attacks to perform directroy traversal and file inclusion. Directroy traversal and file inclusion is when the attackers send command are made to give them access to the directories to put, get and alter files within the computers database. Below are the result of the query which are showing that there where multiple attepts that failed (Failure is shown by status codes 404 and 531), but the results also show that there where two attepts that where succsessful. If this where a proffesional report to a company I, based on the data that is provided would reccomend furthur defenses to be implemented to prenent directory traversal and file inclusion attacks

Directory Traversal and File inclusion

Below are the resluts to the hive query search on a Cross Site Request Forgery which pertains to browsers Javascript alert notice. Attacks such as these are one such attack that uses password phishing in its attack. As one can see there are results in the data below, though if this where a where a proffesional report to a company I, based on the data that is provided would reccomend furthur defenses to be made that prevents Cross Site Request Forgery from happening againm which in turn leaves the computer more secure in the long run.

Cross Site Request Forgery

Below are the results to the hive query that checks the data on possible command injection attacks. These attacks try to disguise commands with HTML URL encoding. As one can see below there are results shown within the data.

Command injection

Below are the results to the hive query that checks the data on possible MySQL Charset Switch and MS-SQL DoS Attack. This attack involver atlering a character set that is used for functionality of databases and avoiding DDOS attack. As can be seen below there are results where the attacker changed the character set gbk_chinese_ci.

MySQL Charset Switch and MS-SQL DoS Attack

Below are the results to the hive query that checks the data on the most failed requests sent to the server. This data would be able to tell us when the most failed requests happen and help us identify the specific host

and day which had the most failed access which could likely be an attacker. The host with the most accesses should be investigated in the future

Hosts with the Most Failed Requests

Below is the results if the hive query which was searching the access logs for bot activity. Below lists the results and the hits in which the bot was found. Possible defenses in the future could include verification of access to the apache server.

Bot Activity

Below are the resultsof the Hosts with the Most Failed Requests per Day and per month. This data can tell us what host to investigate undedr the suspicion of more failed requests, which can be a sign that the host is an attacker trying to send commands.

Hosts with the Most Failed Requests per Day

Hosts with the Most Failed Requests per Month

Below are the data results from hive where the data has been calculated as a ratio of failed to successful requests as a time series. This data can tell us th yeay and the month where the ratio was the highest. These results can tell us what year that should be analysed more in depth. As shown wihtin the results the year and month with the highest ratio Ratio of Failed to Successful Requests as a Time Series

R Code of the Ratio of Failed to Successful Requests as a Time Series

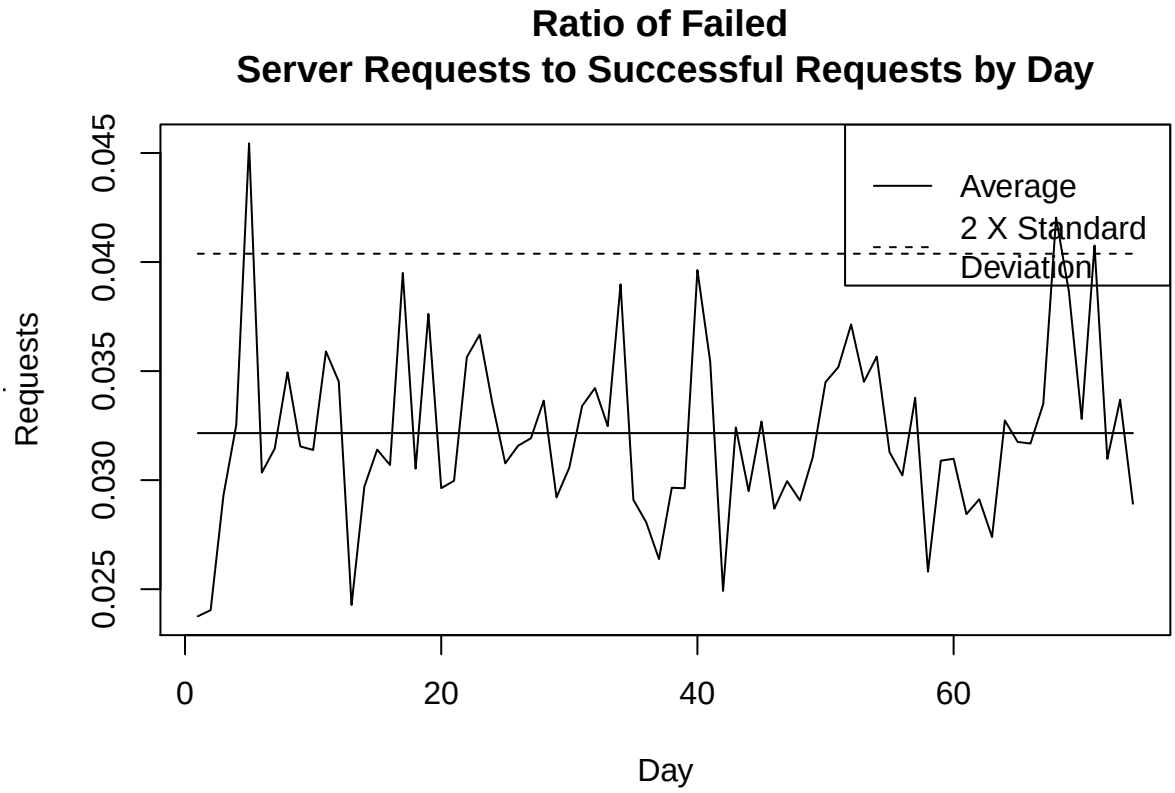
The codeing to the plots below have been taken from chapter 3 of Information Security Analytics: Finding Security Insights, Patterns, and Anomalies in Big Data by Mark Ryan M. Talabis, Robert McPherson, I. Miyamoto and Jason L. Martin

```
library(fBasics)
library(readr)
failedRequests <- read.csv("~/Desktop/untitled folder/failedrequestsbyday.csv")

colnames(failedRequests) <- c("Date","FailedRequestsRatio")
stdev <- sd(failedRequests$FailedRequestsRatio)
avg <- mean(failedRequests$FailedRequestsRatio)
avgPlus2Stdev <- avg + 2 * stdev
failedRequests[failedRequests[,2]>avgPlus2Stdev,]
```

```
##        Date FailedRequestsRatio
## 5  20090724          0.04544236
## 68 20090925          0.04203776
## 71 20090928          0.04075795
```

```
plot(failedRequests[,2],type='l',main="Ratio of Failed
Server Requests to Successful Requests by Day"
,xlab="Day",ylab="Ratio of Failed Requests to Successful
Requests")
lines(rep(avg,length(failedRequests[,2])))
lines(rep(avgPlus2Stdev,length(failedRequests[,2])),lty=2)
legend("topright",c("Average","2 X Standard
Deviation"),lty=c(1,2))
```

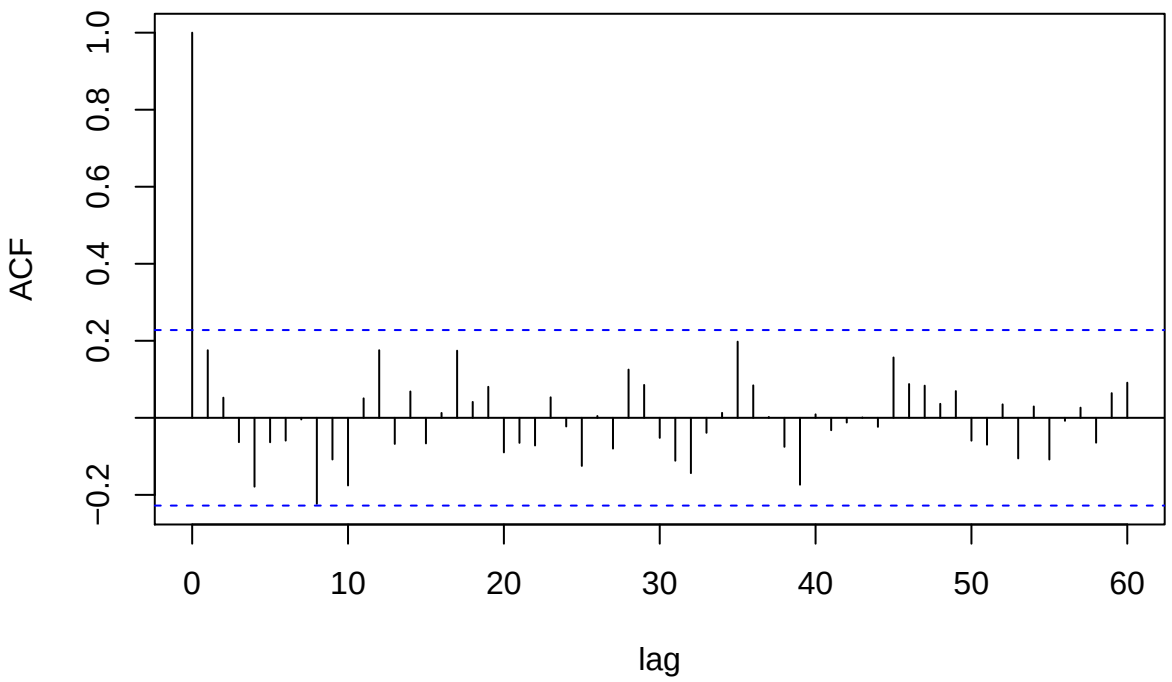


Below is a link to the plot in case the one generated is not to your liking.

Ratio of Failed Server Requests to Successful Requests by Day

Above is a plot of the ratio of failed server requests to succsesful requests by day.It shows the times in which the number of requests have exceeded the threshold average. The three times in which the ratio has exceeded the threshold is July 24, 2009; September 25, 2009; and September 28, 2009. Furthur investigation would be needed on the dates listed to determine if the cause was seasonal or done by other means

```
acfPlot(failedRequests[,2],lag.max=60)
```



Below is a link to the plot in case the one generated is not to your liking.

Ratio of Failed Server Requests to Successful Requests by Day aggregated further

The above plot shows an further aggregation to the data of the ratio of failed server requests to succsesful requests by day.

Conclusion

Based on the provided access logs and the analysis that followed the apache server is quite vulnrable and is in need of various defenses to defend against furthur attacks. I have learned the process in which one can use cloudera and hive to easily sort large amounts of data.

Works Cited

"Log Files." Log Files - Apache HTTP Server. N.p., n.d. Web. 22 Apr. 2017.

Pig Tutorial (n.d.): n. pag. Apache Software Foundation. Web. 22 Apr. 2017.

Talabis, Mark , Robert Mcspherson, I. Miyamoto, and Jason Martin. "Information Security Analytics." Google Books. N.p., n.d. Web. 22 Apr. 2017.