# Data Goals:

See what content packs have been downloaded and by who, how many times.

See what customers are at what server level.

# Data Source:

Logs from CDN (may want to automate the pull on a monthly (or other periodic) basis.

# 20150818

Plan for today is to redo in R, the preprocessing I was doing in Bash via Awk, et al.

Instead of doing everything at once like I’ve been doing, there may be more value in maintaining a separate complete list of solutions and also one for server versions. They can always be merged into a third data set in a separate file.

I may set up a VM with RStudio (or just R) to do the processing for KCL

# Solution Data Harvest:

166.78.9.169

/var/log/nginx . I copy the desired access.log\*.gz files to my chughes/data directory, then use SCP from my local system to retrieve them.

This is the Bash code I’ve been using to preprocess the solutions relevant data.

|  |
| --- |
|  |
| grep "GET /files/published/" access.log | awk '{print $7, $11, $NF}'| grep -v "\"\-\"" | sed 's/\"//g' | sed 's/https:\/\// /g' | sed 's/\// /g' | awk '{print $3","$6","$NF}'| grep -v errata | sort -u > `date +"%Y%m%d\_%H%M%S"`\_solutionmap.csv |
|  |

1. Retrieve lines with “GET /files/published/”
2. Select appropriate columns
3. Discard rows without a requestor
4. Pull out timestamp
5. Pull out Solution name

This is the Bash code used to get the Server Version info.

|  |
| --- |
|  |
| grep v\= access.log | grep -v "\"\-\"" | awk '{print $7, $11, $NF}' | awk -FS"?v=" '{print $1,"\t", $2}' | awk -F'[&]' '{print $1 $2 $3 $NF}' | awk '{print $2 $3 $NF}' | awk '{print $2 $3 $NF}' | sed 's/\=/ /g' | sed 's/\"/ /g' | grep v\= access.log | grep -v "\"\-\"" | awk '{print $7, $11, $NF}' | awk -FS"?v=" '{print $1,"\t", $2}' | awk -F'[&]' '{print $1 $2 $3 $NF}' | awk '{print $2 $3 $NF}' | awk '{print $2 $3 $NF}' | sed 's/\=/ /g' | sed 's/\"/ /g' | sed 's/\// /g' | awk -F\/ '{print $1 $2 $3 $NF}' | sed 's/https://g' | awk '{print $1","$5","$NF}' | grep -v errata | sed 's/s//' | sort -u > `date +"%Y%m%d\_%H%M%S"`\_contentmap.csv |
|  |