MSDS650 Week 4 Tableau Assignment Explanation - Nathan Worsham

First Data Set(s)

On the first data set for this assignment I was interested in continuing to use a data set that I had created for the week 2 EDA assignment. This is because that data set had geo information data (latitude and longitude) that I did not get to work with in week 2 and I was interested in exploring the data using a map. As mentioned in week 2, the data collected is from June 22, 2015 through October 26, 2015 for the website https://agents.pinnacol.com. In this instance I actually had a relevant work question to answer as we had been going through the motions of disabling connections from all other countries in the world except for the US, so getting an idea of who is using our sites would help determine if this was a bad idea or not. Not just wanting to have one map I combined this with another related data set for the website https://online.pinnacol.com. Pinnacol does have a standard www.pinnacol.com site but that does not include login information, in this case I was only interested in login information. The main difference between the two sources is that "online" is a different technology stack so it is logged differently which meant it did not have the success and failed logins that "agents" had as every connection is a status of 302 (redirect).

After getting all of the data plotted which was relatively simple except that I had to drag my geo information out from "Measures" to "Dimensions" to get it to work. I then wanted to exclude the contiguous United States. Tableau made this easy as the map has a selection tool that allows you to draw irregular shapes and then choose to keep or exclude those values. So after getting my original map, I was able to clone the sheet and make the exclusions to the duplicated sheet in order to leave the original alone. At this point I could then change the data visualization to a table and add the dimension "Geo Info" to see where in the world these other connections where coming from. Finally I decided in both cases it made sense to have a zoom in level of connections specifically from Colorado as that is primarily who Pinnacol does business with. So again, cloning the original worksheet and then simply zooming in on Colorado (which thankfully is a square). Thinking that adding population to the map would be interesting, I did so using Tableau's built-in data layer options for maps. Not surprisingly the majority of logins in Colorado correspond to heavier population areas.

Second Data Set

For the second data set I choose to use data from our firewalls from only one day–10/20/2015. This is data that comes from 12 firewalls, though several are active/standby pairs, so the standby doesn't really do or log anything. Regardless it equated to 8,846,563 records! I started with a large overview and then kept filtering until I found interesting events. In IT security this is known as hunting (hpe.com, n.d.), the term comes from military scenarios where the perimeter is assumed to already be breached so "hunt teams" are sent searching for signs of breach. In this case the large overview was a count of events by action taken by the firewalls. These actions are allow, block, or control. Control is an administrative action, so unless something bad was happening you would only expect to see that happen during normal business hours which is the case here. At 12:05 am, there was a spike in traffic, so zooming in on that (which is as simple as drawing a rectangle around the area in question) and filtering the top 10 I was able to find the source of the traffic. In this case it is two DNS servers causing the majority of the spike. I am not sure if that is unusual for that time of day but DNS servers do a lot of external activity, it would need to be compared against other days to see if similar spikes occur.

Next I started using horizontal bar charts often because bar charts are good for comparing values against each other and my labels for them were long and this makes it easier to read. The first bar chart was action per each firewall (dragging action to color to get this effect). This showed as expected the Delta firewall cluster with the majority of traffic. One interesting thing this did point out is that fwcoreftden02 was getting the majority of the traffic from the core firewall cluster as normally 01 would be getting this load. This can be indicative of a failure of 01 but in this case it is just an administrative change that has not yet occurred. Next I decided to look at top sources, top destination, and top destination ports. This proved to be a bit problematic, because while working with Splunk, I am accustomed to group the top n results and then place the remaining into a "Other" group. This does not appear to be a built in function of Tableau. While researching, I found a couple of sites that offered complex formulas to accomplish this, none of which I was able to get to work. Really it would just add a nice touch but wasn't necessary so I decided it was best to just move on. While much of the results were expected a couple items of interest did stand out:

- The server culebra has a high amount of blocks compared to the rest of the sources–this can be normal since this is a proxy server that is open to the outside world and the firewall is blocking all sorts of evil traffic.

- Google's DNS addresses 8.8.8.8 and 8.8.4.4 are some of the top sources–Pinnacol has an internal DNS server that machines should be talking to instead, malware often uses DNS and subsequently common DNS servers such as Google as a first step once a foothold is established on a computer.

So my final page of my Tableau story on this dataset was to further investigate these Google DNS occurrences. Since this data includes the firewall "fwguestlowry" which is traffic for Pinnacol's guest network. This would include things like tablets, phones, etc. This traffic would be normal to use Google DNS as in fact the DHCP server on that network hands that out as the address to use. So filtering out guest network addresses (192.168.2.*), I was left with a graph that showed just one server getting all of the allows which was srv-gst-pear. Turns out that is the DHCP server for that network, so one final filter points out a handful of machines all with internal addresses. Again these computers should not be even attempting to go to Google DNS, it is great that the firewall is stopping the traffic but if it is indeed malware trying to get out then it is likely trying other ways and this is merely a small sign of a bigger picture. As a result I have shared this information with colleagues and we have begun to investigate these computers further to see why they are exhibiting this behavior. So far to date we have been able to explain some of the them because some of them are laptops that are simultaneously connected to our internal network (wired) and connected to our guest network (wireless). This is leading to a policy change of not allowing these laptops to connect to that particular wireless network. But there are still several (nine in fact) on the list that are desktops that do not have wireless NICs that will require digging.

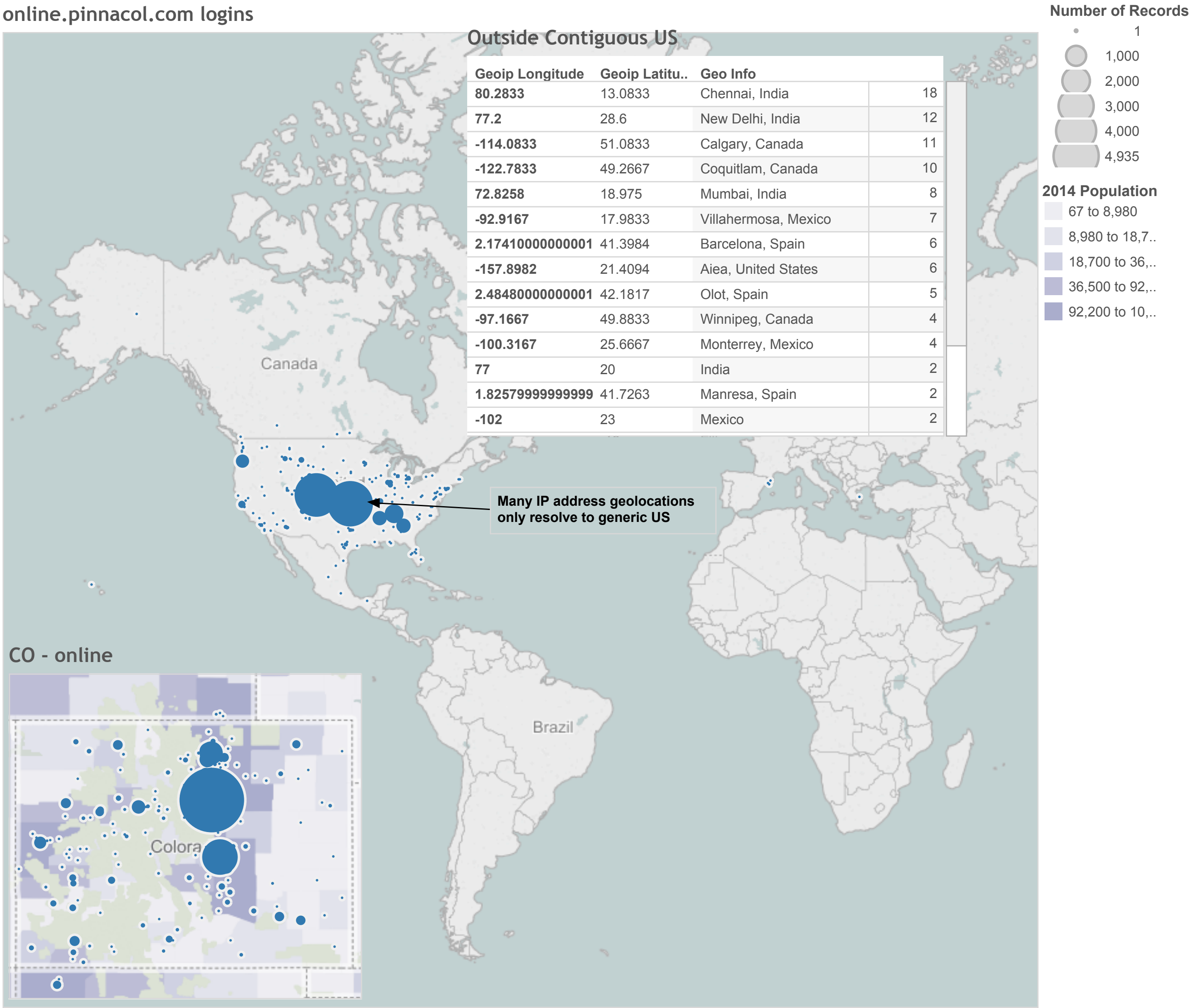
References

hpe.com, n.d. Retrieved from
https://www.hpe.com/h30683/us/en/strategic-business-insights/c/enterprise-security/innovation/how-hunt-teams-can-unmask-hidden-attackers.html

Pinnacol Online Global Activity

Majority of logins occur within US region. Outside the US listed and CO highlighted.

online.pinnacol.com logins



Outside Contiguous US

Geoip Longitude	Geoip Latitu..	Geo Info	
80.2833	13.0833	Chennai, India	18
77.2	28.6	New Delhi, India	12
-114.0833	51.0833	Calgary, Canada	11
-122.7833	49.2667	Coquitlam, Canada	10
72.8258	18.975	Mumbai, India	8
-92.9167	17.9833	Villahermosa, Mexico	7
2.17410000000001	41.3984	Barcelona, Spain	6
-157.8982	21.4094	Aiea, United States	6
2.48480000000001	42.1817	Olot, Spain	5
-97.1667	49.8833	Winnipeg, Canada	4
-100.3167	25.6667	Monterrey, Mexico	4
77	20	India	2
1.82579999999999	41.7263	Manresa, Spain	2
-102	23	Mexico	2

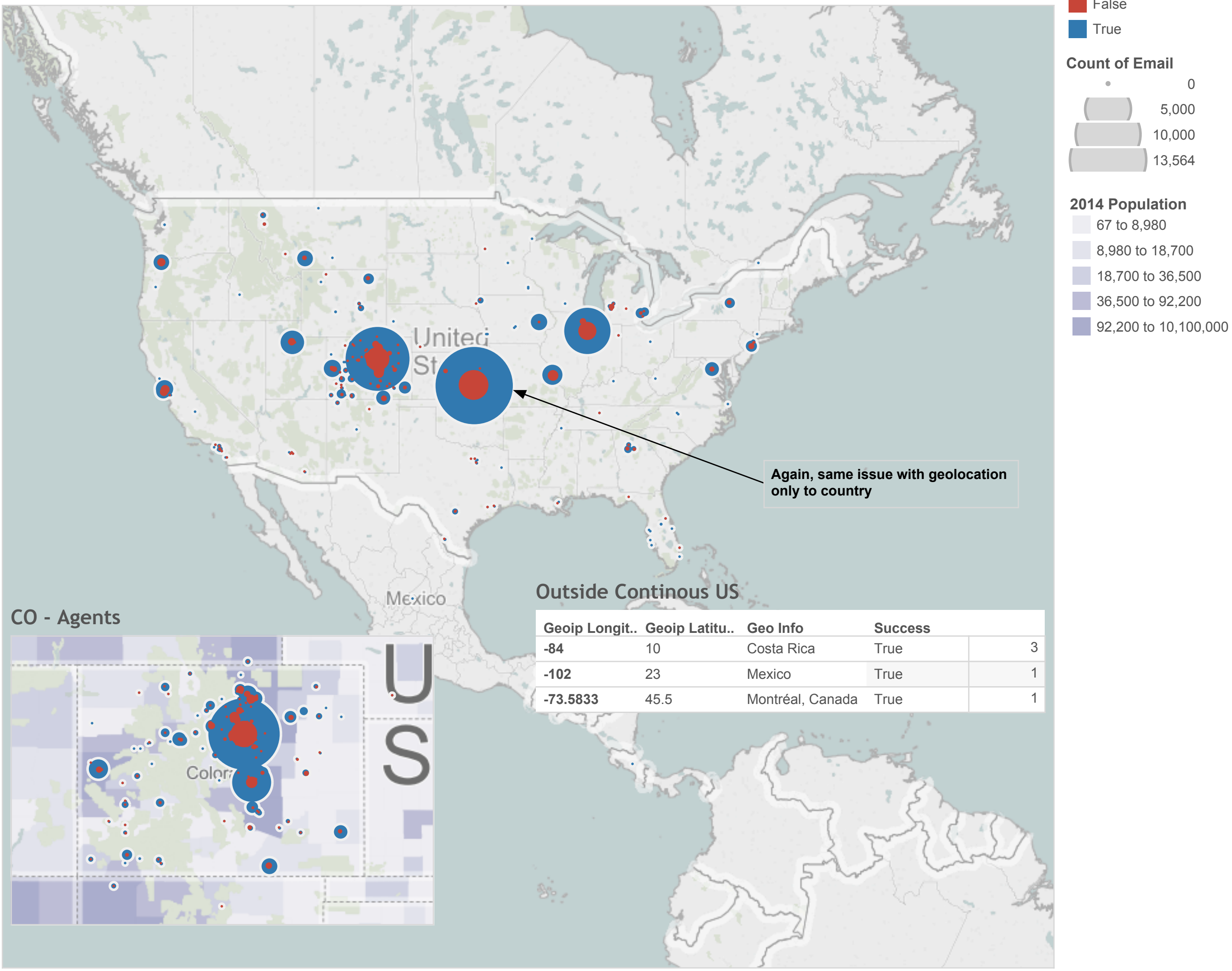
Many IP address geolocations only resolve to generic US

CO - online

Pinnacol Online Global Activity

Traffic is much more clustered around Colorado on agents.pinnacol.com with very few external countries.

agents.pinnacol.com logins (with failures)



Again, same issue with geolocation only to country

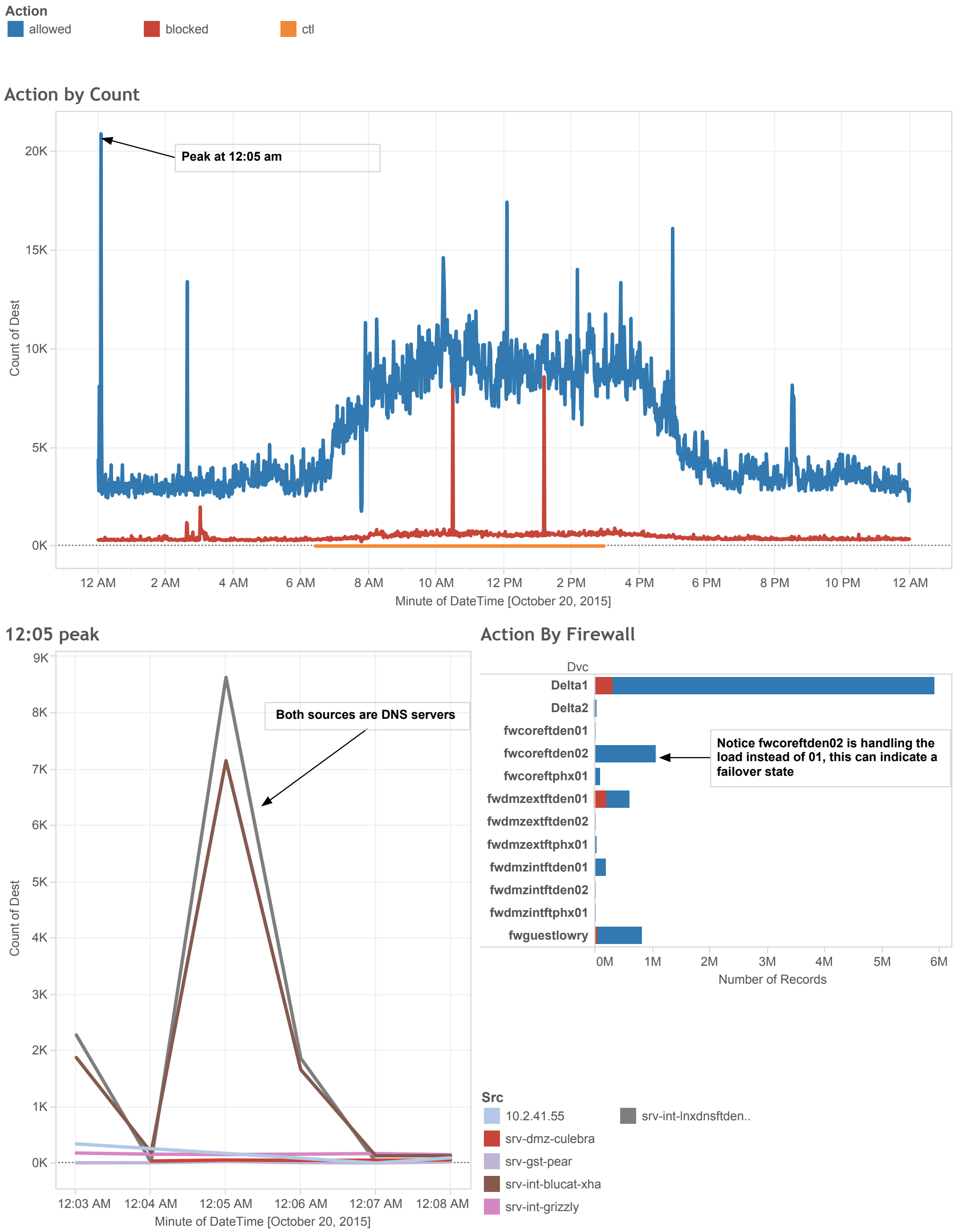
CO - Agents

Outside Continous US

Geoip Longit..	Geoip Latitu..	Geo Info	Success	
-84	10	Costa Rica	True	3
-102	23	Mexico	True	1
-73.5833	45.5	Montréal, Canada	True	1

Firewall Activity 10/20/2015

Firewall activity for 10/20/2015 - 8,846,563 records. High number of allows around 12 am, otherwise follows buisiness hours. "Delta" firewall cluster sees the majority of the traffic.



Action by Count

Peak at 12:05 am

12:05 peak

Both sources are DNS servers

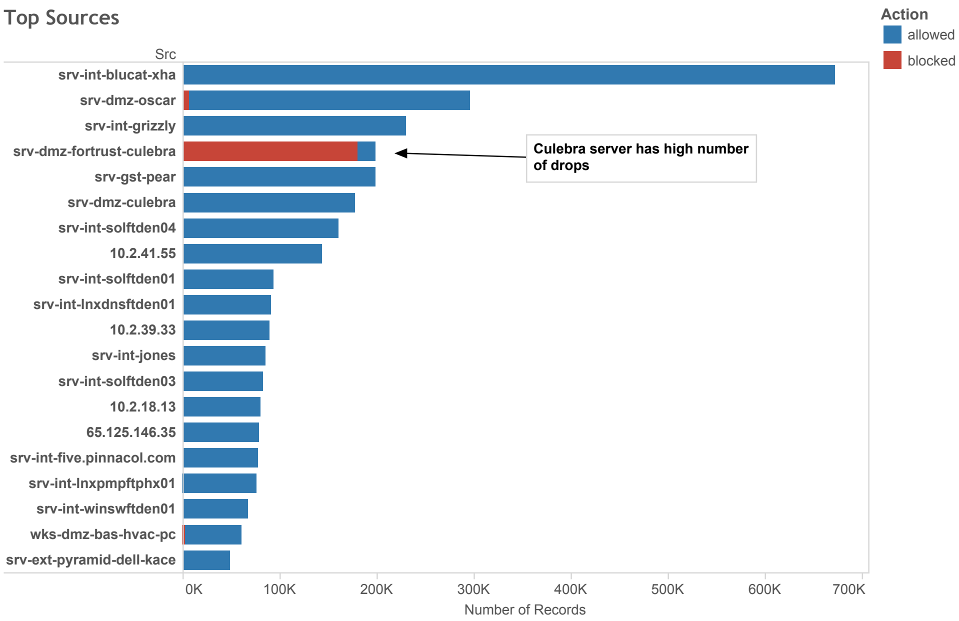
Action By Firewall

Notice fwcoreftden02 is handling the load instead of 01, this can indicate a failover state

Firewall Activity 10/20/2015

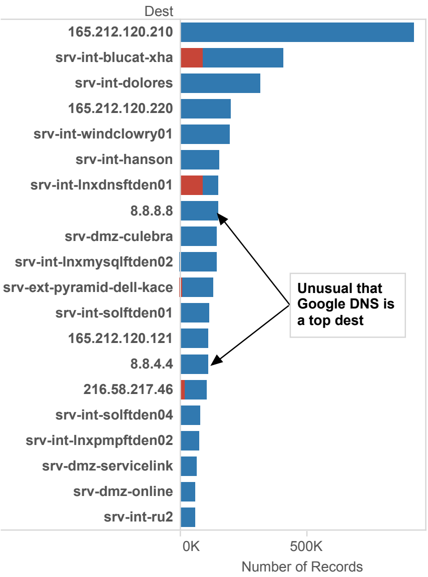
Top sources are unsurprisingly servers with external web functions, top is a DNS server. However there is unusal activity going to Google DNS servers when we look at top destinations.

Top Sources



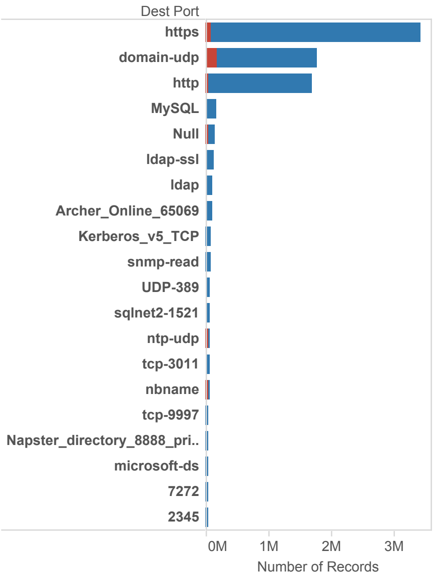
Culebra server has high number of drops

Top Destinations



Unusual that Google DNS is a top dest

Top Destination Ports



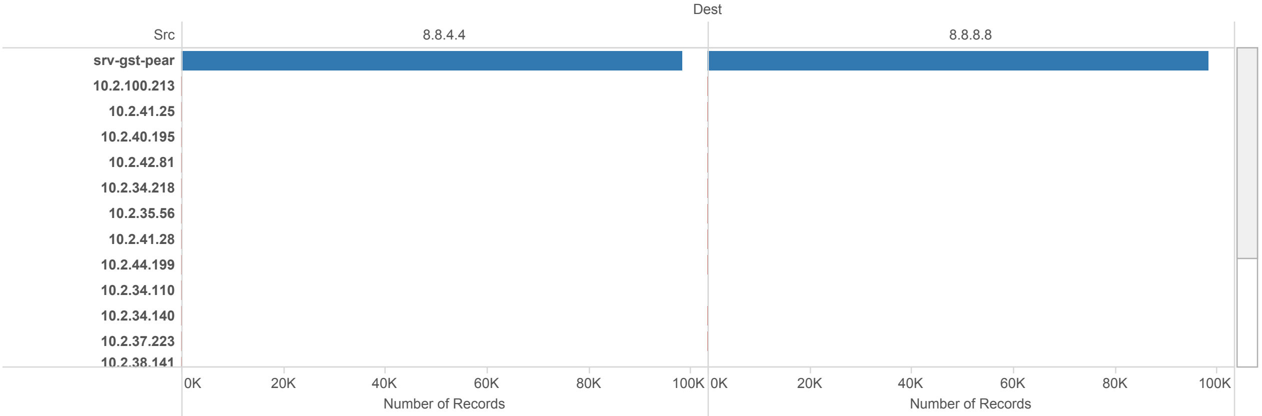
Firewall Activity 10/20/2015

Investigating Google DNS by first filtering out guest network connections show one server with the majority of the connections--srv-gst-pear--that is expected behaviour. Once that is filtered out we see a handful of suspect internal machines.

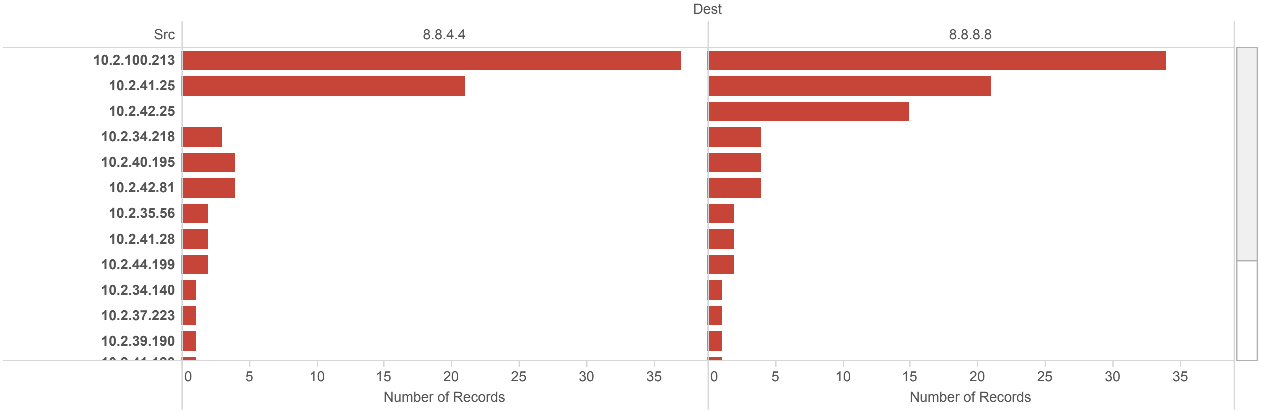
Action

🟦 allowed 🟥 blocked

Google DNS No Guest Network



Google DNS No srv-gst-pear


Dest
Src
Number of Records
Number of Records