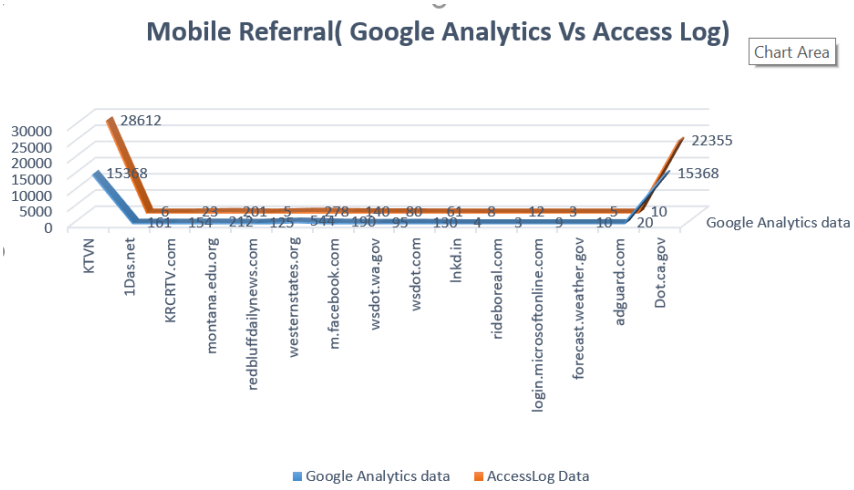
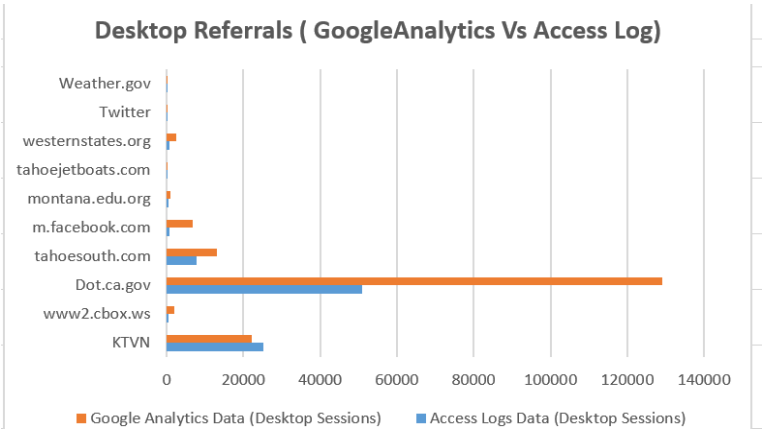
Aayush Goyal
A02259952



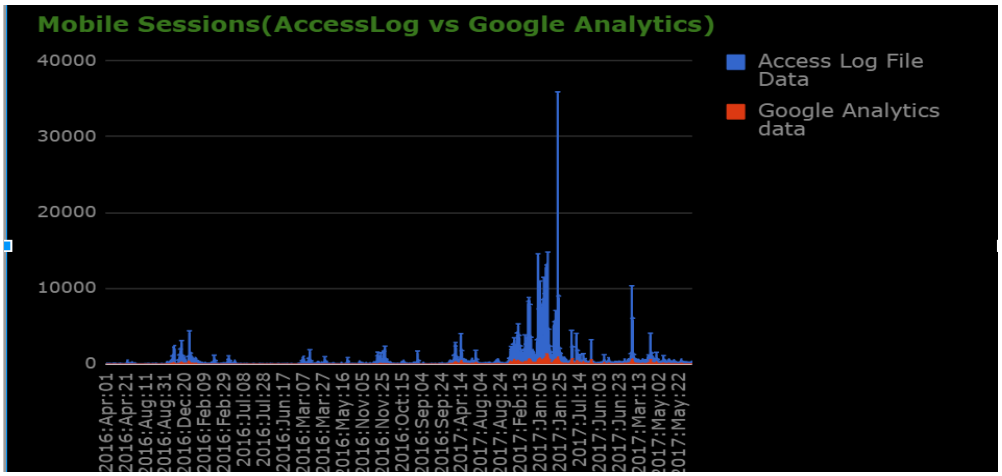
Google Analytics showed the data for the year 2016-17. So I Extracted exactly those year's data from the access.log for some mobile referrals. As we can see from the chart that are some discrepancies between access.log & Google analytics. The reason for the same can be that some sites represent fake Referrals(Ex: Cabelas.com). They are created in the Google Analytics account to trick us into visiting Spammy websites. If we open one of these URLs in the browser, we will likely be redirected to an online store, marketing scam or malware site. This is what screws up the Google Analytics matrics. On the Otherhand log files are literally raw files which need initial processing to be used for web site analytics.



These charts are indicating more referral data in Google Analytics than Log file.

Log files can determine the referring site for a visit most of the time. It is stored with the first page of their visit. This referrer can then be further analyzed to tell if it was a website, a paid advertisement or a search originating from a search engine. A strategy on the Internet that malicious webmasters use is to try and publicize their website by engaging in referrer spam. Which means is that they will create an automated web crawler much like Google's Googlebot. They then send this crawler out to visit hundreds of millions of websites and the crawler pretends to be a normal web browser. The user string that the crawler sends is Chrome or Internet Explorer or another human looking web browser. But the vital alteration is that it sends a fake referrer string to any website it visits. That means that it tells the logging application on any website it visits that it arrived from a website that it didn't actually arrive from.

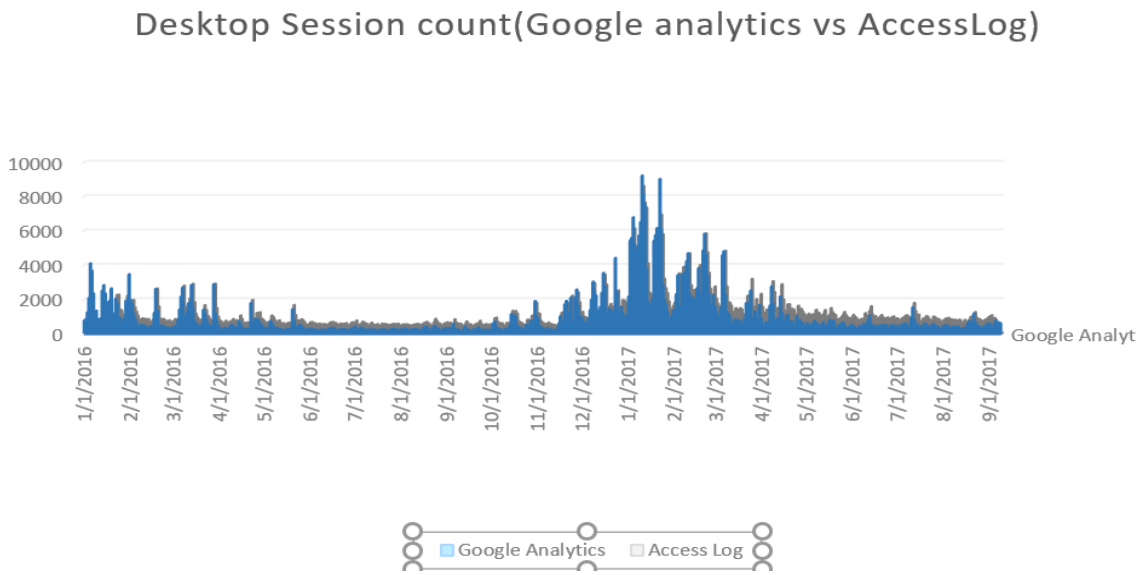
Aayush Goyal
A02259952



In this chart of Mobile sessions between Google Analytics & AccessLog it is quite evident that there are more sessions indicated by AccessLog. An explanation for the same can be the basis on which google analytics & Log files maintain their data.

Mobile & Desktop sessions created in AccessLog are quite large as expected because if our website receives 10,000 visitors a day and each visitor or bot views an average of ten pages, then the server will generate 100,000 log entries within a single log file every single day. Each & every activity like image downloads , clicks etc are tracked by log files whereas in Google analytics it only maintains the html content session.

Log files log requests to all files and media requested by the client. On the other hand analytics are only executed for html content. One must consider that browsers do keep a cached version of the page to their local caches. Thus sometimes when a browser reloads a page (like when the user presses the back button) the browser will most probably load the page from its local cache other that requesting it from the server. Hence for this the call does not get logged.



However analytics records even cached pageviews because the tracking cookie executes every time a page is displayed. This obviously leads to duplicate entries and additional page views. It is

very common for somebody to refresh the page. Because of these variables, the number of total visits and pageviews will never match between analytics and log files.

Conclusion:

By comparing the charts between both Google Analytics & Access.log it can be concluded that both are not wrong at their place, they are exactly doing what they are programmed or designed to do. So if Log files and Google Analytics showed the same thing, one or the other of them would not exist. The reason for discrepancies between the two depends on the metrics the analytics are using which are different from the metrics being used to analyze the log files.

There are some Pros & Cons of both the logging methods like Google Analytics tracks its users by when a user loads his JavaScript into browser. If some users accessed a website with JavaScript disabled, then Google Analytics may not keep their track. So basically If we rely only on Google Analytics or any other platform that relies on front-end scripts, then we see only part of the story as a result of the increasing use of script blockers. To analyze complete server and website activity properly, we need to monitor the server log files directly.

Appendix

Assumptions I took

1. The time required to end a session for the user is 20 minutes.
2. For the Desktop & Mobile data IP address may not be unique so I even considered the browser they were using.
3. According to the session gap of 20 minutes I calculated the time duration when a particular user was active and then divided those minutes by 20 from which I got the total number of sessions that user made.

Approach:

First I separated the data for the years 2016-17 from Access.log because the data given in Google Analytics file was from 2016-17 so for the efficient comparison. I used a combination of command prompt & Intellij idea to process the log files. I created the python script which splitted the whole string present in access log into just those which were useful to me like date, sessions & IP addresses and initialized date library to calculate the time difference to calculate no. of sessions created by IP.Then I segregated Mobile data and Desktop data from that so that I can get the different views of sessions & referrals. Then finally I extracted session data and referral data from both.