DATA EXPLORATION REPORT

Hotel TULIP Weblog Dataset

Abstract

This report comprises of exploratory data analysis for the Hotel TULIP Weblog dataset and representing some numeric insights and visualizations which will help Hotel TULIP to improve their web portal services for better customer experience.

Submitted to

Team 'SIT742'

Prepared By

Mohitkumar Rangholiya (215410048)

Data Analyst, Hotel TULIP

Index

INTRODUCTION 2

HANDLING LARGE DATASETS (HTWEBLOG) 2

EXPLORING INDIVIDUAL FIELDS FROM LOGS 3

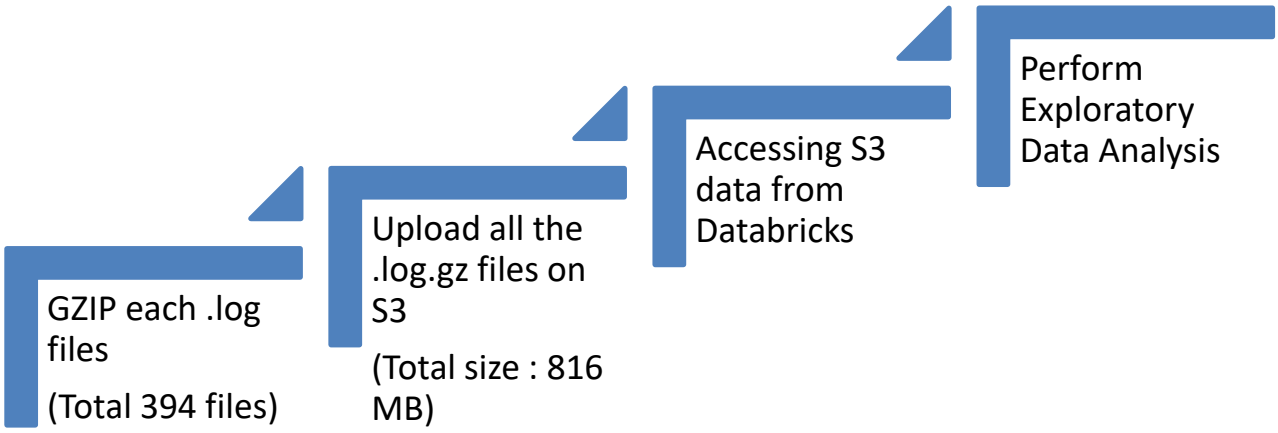
SERVER STATUS CODE COUNT AND VISUALIZATION 3
WINDOWS STATUS CODE COUNT AND VISUALIZATION 4
EXPLORATION OF TIME TAKEN BY SERVER TO RESPOND THE USER REQUESTS 5
EXPLORATION OF REQUEST METHOD FOR ACTION TO BE PERFORMED 7
EXPLORATION OF DATE & TIME AND ITS VISUALIZATION 8
FREQUENT VISITORS OF TULIP SERVER 10
CS_USER_AGENT COUNT 10
TOP TEN CS_URI_STEM 11
PLOTTING FREQUENT PAGES MOSTLY VISITED FROM USERS 12
TOTAL NUMBER OF UNIQUE CS_IP(CLIENTS) VISITED TULIP WEBSITE 12
SLOWEST 20 URIS 13

Introduction

Here the Hotel TULIP server access log data is provided to explore and find some findings from that which can help our Hotel to improve the web services and level of customer satisfaction and experience. So firstly as a data analyst from TULIP Hotel IT Division, my task is to load the Bigdata of logs generated by server and then explore it to find some interesting insights.

Handling Large Datasets (HTWebLog)

TULIP server logs are taken from August,2014 to September,2015 and hence the size of the data is about 15.8 GB. Here is how I managed to load whole data on Amazon S3 and access it from Databricks and then explore it.

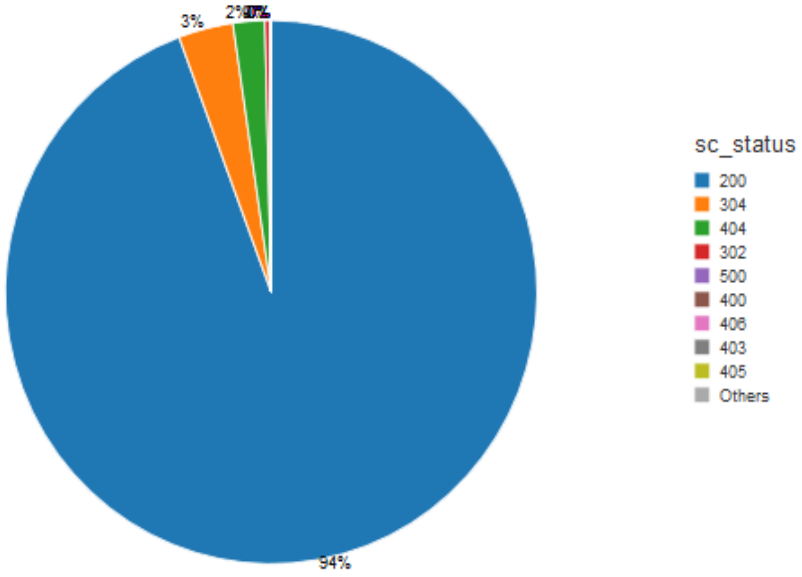


Exploring individual fields from logs

Server status code count and visualization

- Below table shows the total number sc_status categories and occurrence of each sc_status codes during 13 months period (August,2014 to September,2015).
- It is clearly seen from the table that '200' status code is much more larger than all other (406,403,405,301,501) status codes which is highlighted using green and red color respectively.
- The given pie chart reflects the visualization for the table demonstrated here. So from pie chart we can say that **94% of the user-requests have been responded correctly without error.**

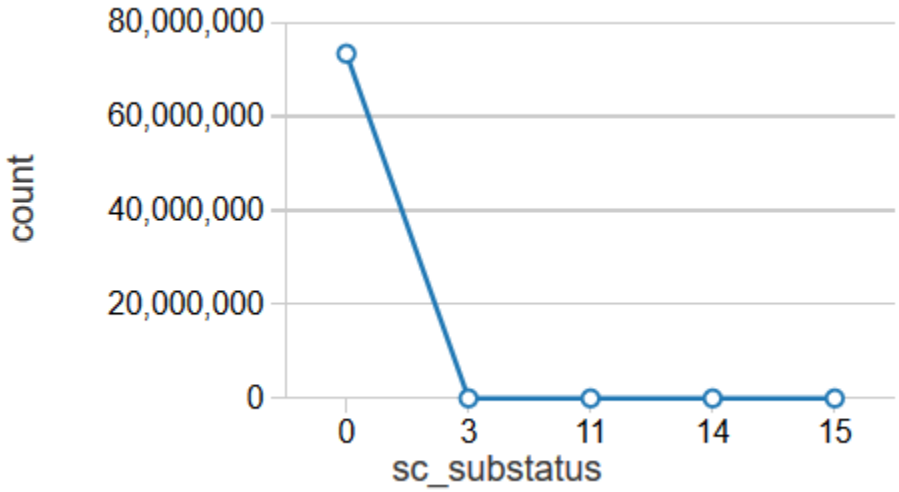
sc_status	count
200	69236725
304	2443217
404	1408877
302	237559
500	35490
400	5236
406	379
403	319
405	273
301	180
501	1



Server sub-status code count and visualization

- Similarly we can count and plot server sub status as given below.
- Here it is noticeable that almost 100% of the records are having “0” substatus. At the same time other types of substatus codes are negligible and hence the line graph represents almost zero for others(3,11,14,15).

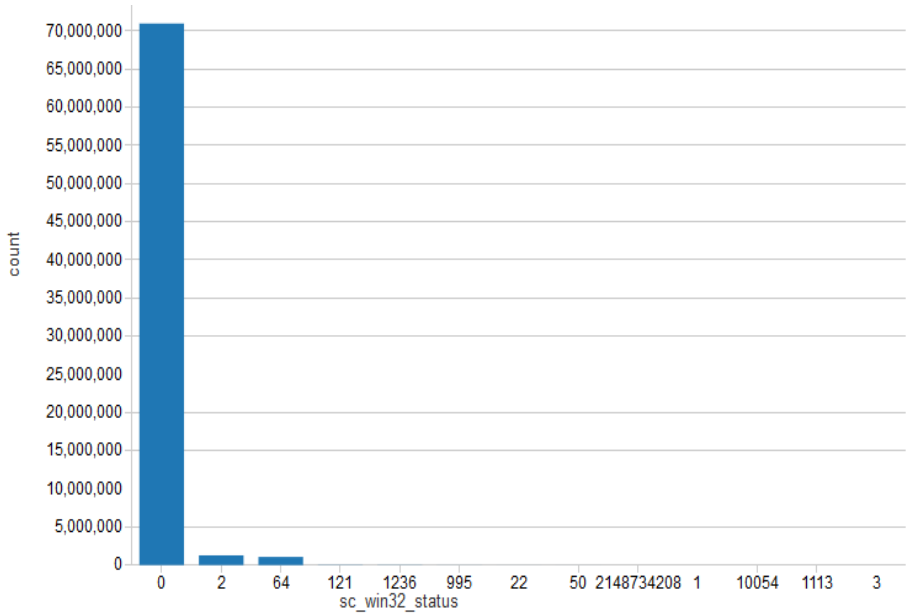
sc_substatus	count
0	73364103
3	2980
11	847
14	319
15	7



Windows status code count and visualization

- Below table representation depicts the total number of sc_win32_status categories and occurrence of each value during 13 months period (August,2014 to September,2015).
- It is clearly seen from the table that ‘0’ status code is much more larger than all other windows status codes which is highlighted using green color.
- The given bar chart reflects the visualization for the table demonstrated here. So from bar chart we can see the variations. There are **70951439 of the user-requests have been fulfilled successfully.**

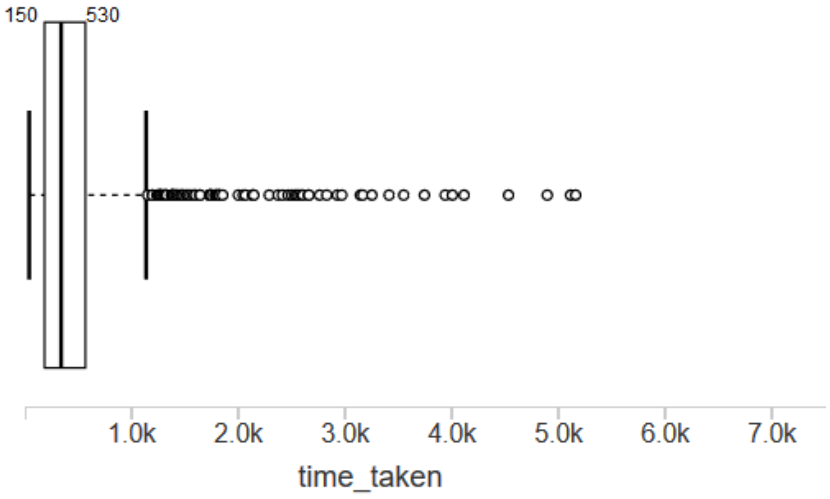
sc_win32_status	count
0	70951439
2	1276889
64	1064653
121	33758
1236	25486
995	9640
22	2969
50	2851
1	269
10054	17
1113	3
3	1



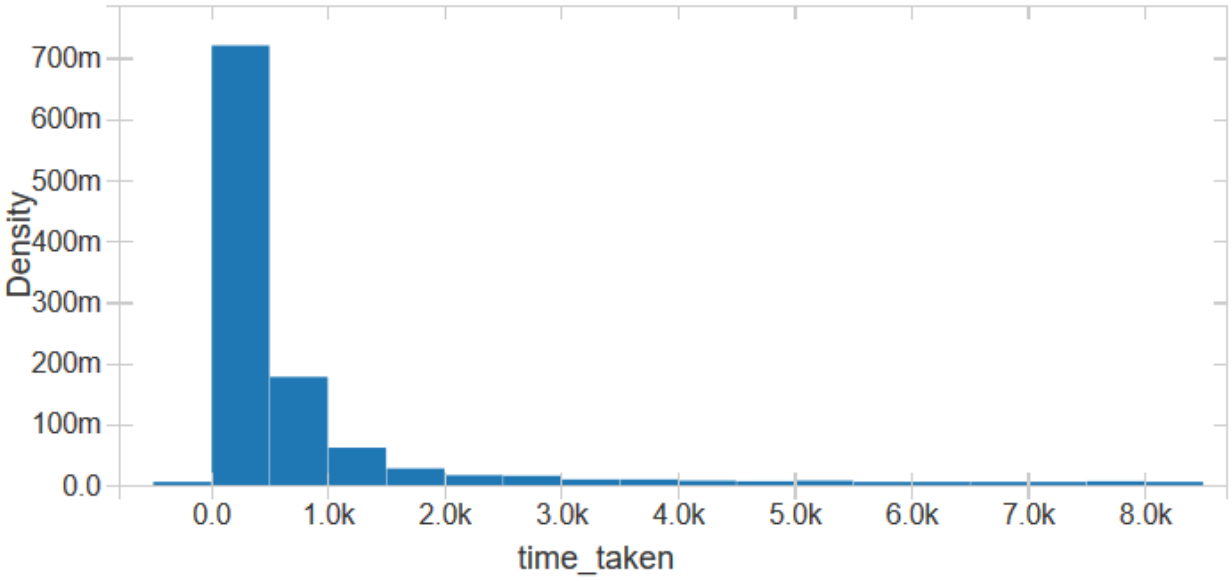
Exploration of time taken by server to respond the user requests

- Time taken to satisfy user request is given in milliseconds. So we can easily calculate summary statistics for that and then plotting boxplot is a better way of visualizing it.

summary	time_taken
count	73145699
mean	620.2576000811749
stddev	4903.5757797471615
min	0
max	8238493



- Histogram representing the time taken using the bucket size of 1000 milliseconds.

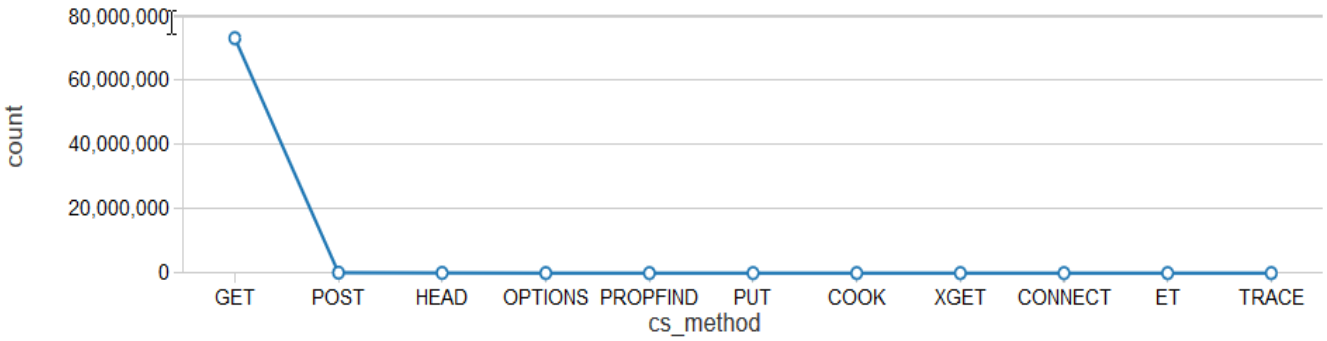


Exploration of Request method for action to be performed

- Cs_method is defined by client browser and this is the method of request in which the response will take place. Below tabular form depicts that GET method is the highest requested method from all the requests to the server.

cs_method	count
GET	73161951
POST	141028
HEAD	60943
OPTIONS	2563
PROPFIND	1046
PUT	563
COOK	158
XGET	1
CONNECT	1
ET	1
TRACE	1

- Here, provided pie chart and line graph represents that GET method is almost 100 % because other methods are very minor as compare to GET which can be easily seen from line graph provided below.

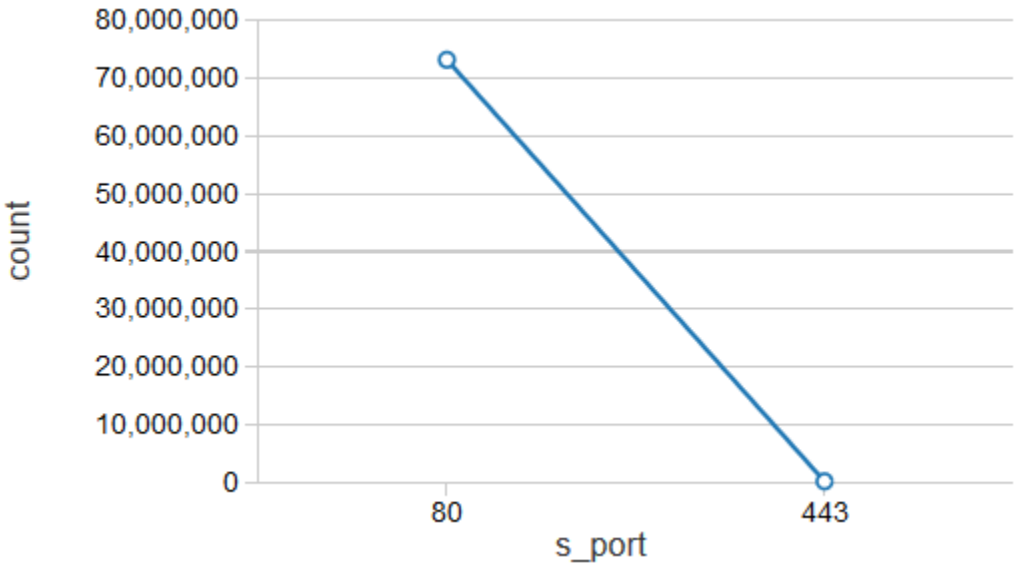


Exploration of Server Port and its visualization

- In the Server port field , there are only two types of values as per described in the given table.

s_port	count
80	73220811
443	147445

- Line graph describing the count of server port for '80' and '443'. It dropped significantly as there is a huge difference between their counts.



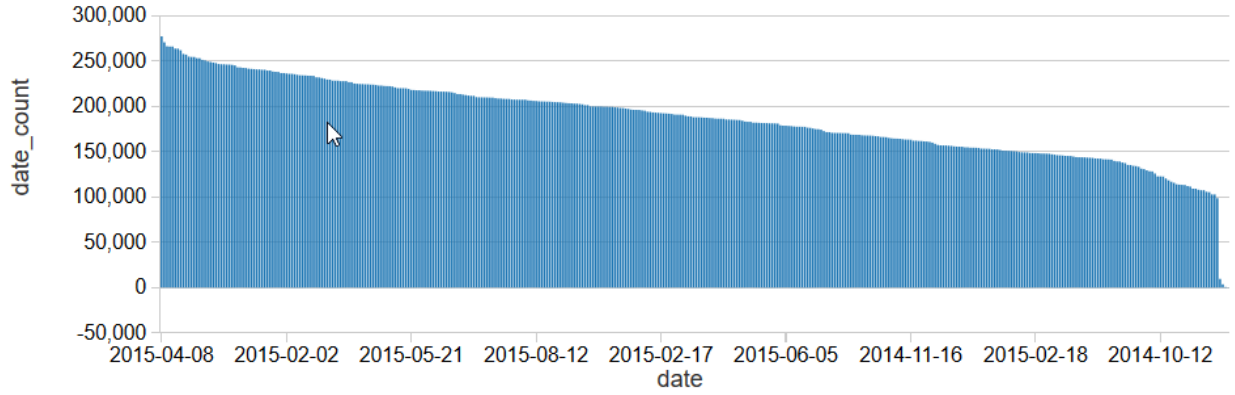
Exploration of Date & Time and its visualization

- Top 10 busy days when TULIP server has got nearly 70% more requests than average requests(**186213**) in whole duration of 13 months.

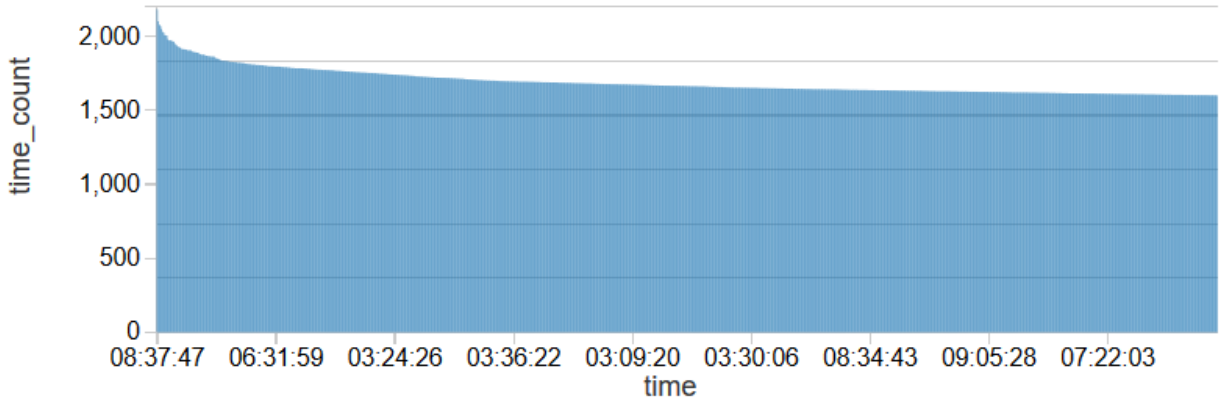
date	date_count
4/8/2015	277331
3/18/2015	270800
3/23/2015	266422
4/23/2015	266201
4/28/2015	266176
1/27/2015	263841
12/2/2014	263776
4/16/2015	262003
3/24/2015	257945
1/26/2015	257184

date_count	
Mean	**186213**.8477
Standard Error	2161.523556
Median	187986
Mode	#N/A
Standard Deviation	42905.01752
Sample Variance	1840840528
Kurtosis	1.920658887
Skewness	-0.754513454
Range	277071
Minimum	260
Maximum	277331
Sum	73368256
Count	394

- Bar graph describing the date wise total count of requests to the servers.



- Bar graph describing the time wise total count of requests to the servers. It is clearly seen from the graph that the TULIP server has got the **highest requests during early morning hours**.



Frequent visitors of TULIP Server

- Top 10 cs_ip that have accessed server more than 100 times.

```
Any 10 cs_ip that have accessed more then 100 times: [u'165.212.246.60', u
'93.33.66.143', u'117.19.215.93', u'203.198.14.88', u'107.77.68.78', u'92.
149.180.130', u'92.40.249.41', u'123.203.164.95', u'203.186.99.98', u'70.5
0.232.11']
Command took 15.62 minutes -- by mranghol@deakin.edu.au at 4/14/2017, 10:44:08 PM on
sit742
```

Cs_user_agent count

- TULIP server has got highest number of requests from Mozilla browser version 5.0 from Windows OS. Likewise, we can refer other top user_agents from this table.
- Word Cloud can be generated from this String of cs_User_Agent.

cs_User_Agent	count
Mozilla/5.0+(Windows+NT+6.1;+WOW64;+Trident/7.0;+rv:11.0)+like+Gecko	3695427
Mozilla/5.0+(Windows+NT+6.1;+Trident/7.0;+rv:11.0)+like+Gecko	1628733
Mozilla/5.0+(iPhone;+CPU+iPhone+OS+8_3+like+Mac+OS+X)+AppleWebKit/600.1.4+(KHTML,+like+Gecko)+Version/8.0+Mobile/12F70+Safari/600.1.4	1288186
Mozilla/5.0+(compatible;+MSIE+9.0;+Windows+NT+6.1;+Trident/5.0)	1270461
Mozilla/5.0+(compatible;+MSIE+9.0;+Windows+NT+6.1;+WOW64;+Trident/5.0)	1206858
Mozilla/5.0+(iPad;+CPU+OS+7_1_2+like+Mac+OS+X)+AppleWebKit/537.51.2+(KHTML,+like+Gecko)+Version/7.0+Mobile/11D257+Safari/9537.53	865130
Mozilla/5.0+(iPhone;+CPU+iPhone+OS+8_1_2+like+Mac+OS+X)+AppleWebKit/600.1.4+(KHTML,+like+Gecko)+Version/8.0+Mobile/12B440+Safari/600.1.4	786734
Mozilla/5.0+(compatible;+MSIE+10.0;+Windows+NT+6.1;+WOW64;+Trident/6.0)	649436
Mozilla/5.0+(iPhone;+CPU+iPhone+OS+8_1_3+like+Mac+OS+X)+AppleWebKit/600.1.4+(KHTML,+like+Gecko)+Version/8.0+Mobile/12B466+Safari/600.1.4	641430
Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/39.0.2171.95+Safari/537.36	640924

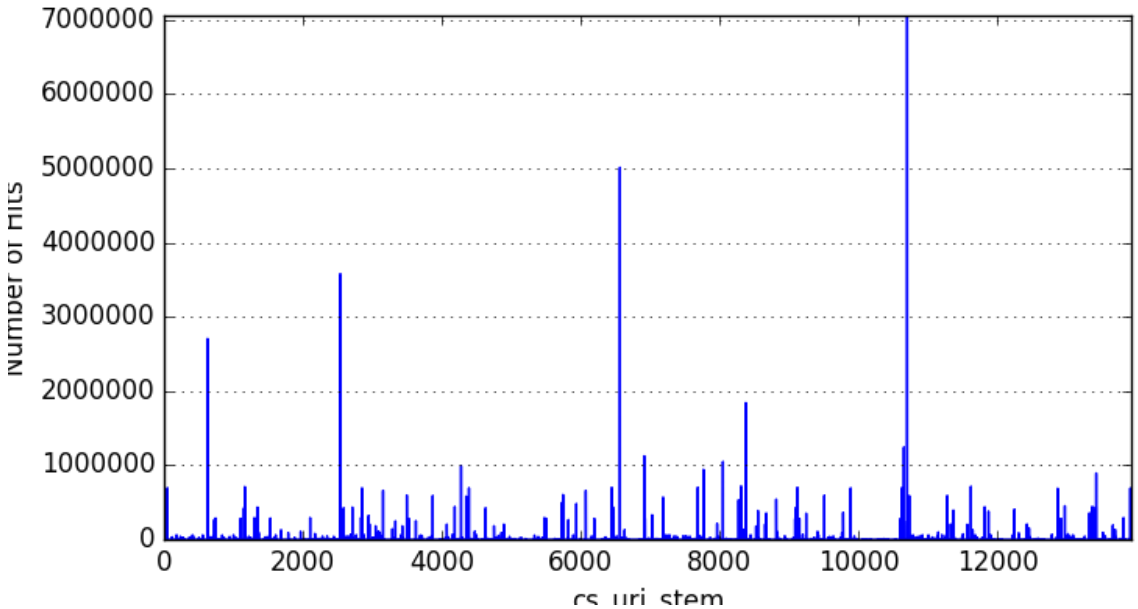
only showing top 10 rows

Top Ten cs_uri_stem

No	cs_uri_query	hits
1	/~/media/Images/Hotel_ICON_revamp/about+us/icon/sitecore_media.ashx	7083000
2	/~/media/Images/Hotel_ICON_revamp/home/sitecore_media.ashx	5028816
3	/layouts/Layouts/Hotel_ICON_revamp.aspx	3593303
4	/~/media/Images/Hotel_ICON_revamp/dining/large/sitecore_media.ashx	2715378
5	/~/media/Images/Hotel_ICON_revamp/offers/sitecore_media.ashx	1845107
6	/~/media/Images/Hotel_ICON_revamp/rooms/Above+and+Beyond/sitecore_media.ashx	1248132

7	/~/media/Images/Hotel_ICON_revamp/Navigation+Bar/facilities/sitecore_media.ashx	1130114
8	/~/media/Images/Hotel_ICON_revamp/Navigation+Bar/Events/sitecore_media.ashx	1054804
9	/~/media/Images/Hotel_ICON_revamp/Navigation+Bar/Rooms/sitecore_media.ashx	996203
10	/~/media/Images/Hotel_ICON_revamp/about+us/award/sitecore_media.ashx	949540

Plotting Frequent pages mostly visited from users



Total number of Unique cs_ip(clients) visited TULIP website

- Unique cs_ip directly reference to the **number of visitors** of TULIP Web portal during given period of time. So we can refer it that total “522801” visitors visited the TULIP Web portal during given 13 months.

```
Unique cs_ip: 522801

Command took 17.57 minutes
```

Slowest 20 URIs

- Find the Slowest 20 URIs (in average time taken) while visiting TULIP web-portal.
- It is sorted by the average time taken by specific cs_uri_stem.

cs_uri_stem	Max	Min	Average
/~/media/Files/Hotel_ICON_revamp/pdf/Career/sitecore_media.ashx/trackback/	3744569	20850	966398.1967213114
/~/media/files/hotel_icon_revamp/pdf/career/sitecore_media.ashx	211511	4416	68537.81818181818
/~/media/Files/Hotel_ICON_revamp/pdf/Career/sitecore_media.ashx	3514480	31	60702.52135306554
/careers-and-education/~/media/Files/Hotel_ICON_revamp/pdf/Career/sitecore_media.ashx	120191	176	36614.5
/~/media/Files/Hotel_ICON_revamp/pdf/Career/1407+-+SHTM+Research+Project.ashx	623100	31	32843.1724137931
/about-the-hotel.aspx/RK=0/RS=CnBWyiCOOSGRDhtZEmHuSk1ZJ0s-	26014	26014	26014.0
/images/our-city/default_no_event_Dec.jpg	118535	178	21395.894736842107
/~/media/Files/Hotel_iCOn_revamp/pdf/Career/sitecore_media.ashx	19378	19378	19378.0
/m/~/media/Files/Hotel_ICON_revamp/pdf/Career/sitecore_media.ashx	19279	19279	19279.0
/Wine_and_Dine/What_is_On/Peruvian_Food_Festival.aspx/	16476	16476	16476.0
/vip.rar	16448	16448	16448.0
/r/www/cache/static/home/img/logos/nuomi_ade5465d.png	20491	12723	16088.0
/offers.aspx/trackback/	25453	5669	15561.0
/zc/chs/img/body.png	19911	11784	15276.777777777777
/~/media/files/hotel_icon_revamp/pdf/events/silverbox-ballroom/silverbox_ballroom_floorplan.ashx	18396	11364	14880.0
/Press.aspx/RK=0	23452	6021	14736.5
/~/media/5104d65aa9d448ccaf8595214e021aa6.ashx	14707	12117	13412.0
/en/css/js/jquery/jquery.js	25207	52	12629.5
/~/media/Files/Hotel_ICON_revamp/pdf/offers/201407+-+Summer+Meeting.ashx	1768946	41	12230.61797752809
/hk/	11745	11745	11745.0

only showing top 20 rows