Data analytics report for the Hotel TULIP Web log dataset.

Python Spark

MOHIT RANGHOLIYA

Contents

Background & Objectives

The team- SIT742 has got the exploratory data analysis report from the Hotel TULIP's IT division. After going through the exploratory results, our team has decided the objectives of this analysis and the flow of the analysis to be proceeded using full Hotel TULIP Web log dataset.

Data doesn't contain information about sessions. While exploring data, we came across multiple entries having accessed the server in a single login.

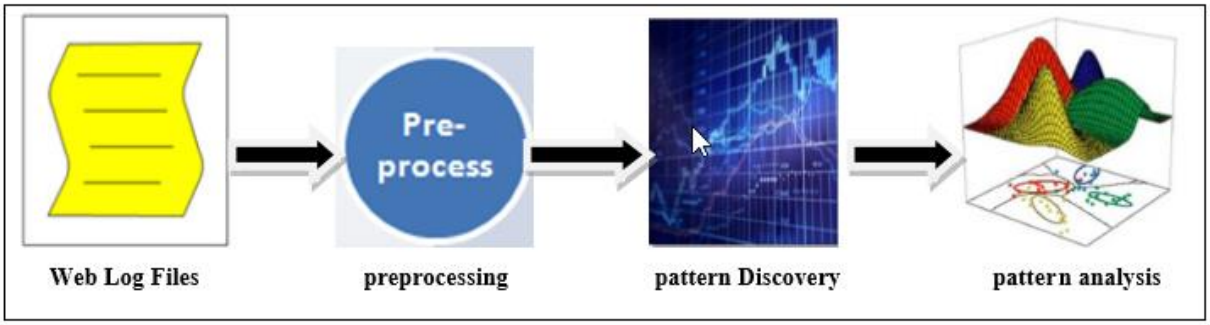
Identification of the session ID

To carry out effective analysis, our first objective is to identify the session ID in the web log files. As the web log files are in frequency of daily, we selected once attribute as the date. To identify the user, we consider the client IP as the second attribute. In addition, to make the session ID unique we selected the user agent as the third attribute which contains browser details. Finally, the session ID constructed based on concatenation of date, client IP and user agent attributes.

User's frequent access pattern analysis

Our second objective is to analyze the webpage-patterns for user's server access using the session data based on identified session Ids. We believe it will help to find useful insights on user web page access patterns.

Process-flow



As per described in the above diagram(CU & Bhargavi 2013), the standard process has been followed to carry out the whole analysis.

Data Preprocessing

It was clear that we need to consider only the required features and filter unnecessary/invalid data to align with our objectives.

As we are interest on web pages to identify the user patterns, we consider only the name of web pages presented in a given web log record. To identify this web page, we manipulate web page name by using a user defined function on uri stem attribute as below.

```
# Load the space-delimited web logs (log files) and assign to column names.
row_data = split_data.map(lambda p:Row(

                          # formulate date_time_stamp
                          date_time = datetime.strptime(p[0]+" "+p[1],"%Y-%m-%d %H:%M:%S"),

                          # assign row data from log file
                          date = p[0],
                          time = p[1],
                          cs_uri_stem = p[4],
                          cs_ip = p[8],
                          cs_user_agent = p[9],
                          sc_status = int(p[10]),
                          sc_substatus = int(p[11]),
                          sc_win32_status = int(p[12]),
                          time_taken = int(p[13]),
                          session_Id = p[0]+p[8]+p[9],
                          path = generate_sections_of_url(p[4])
                          )
                          )
```

Feature selection & Data Reduction

Feature Selection

As, per the data dictionary, we have total 14 features in the log files generated by TULIP Server. There are some of the features that we decide to ignore based on our objectives and data exploratory analysis on earlier stage. Hence, we dropped below six features from the web log data set.

Attribute Name
s_ip
cs_method
cs_uri_query
s_port
cs_username
s_win32_status

We have selected below 8 features, which are significance for further analysis based on our objectives.

Attribute Name
date
time
cs_ip
cs_uri_stem
user_agent
time_taken
sc_status
sc_substatus

Data Reduction

Features considered to identify sessions are:

Attribute Name
date
cs_ip
user_agent

Sample session ID shows as below; before use the has function.

'2014-08-
021.174.103.239Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chrome/36.0.1985.125+Safari/537.36'

During our initial data exploratory stage, we identified web log records with blank data for user agent attribute. Hence, whenever the user agent is blank ("-"), we dropped the web log record associated, as session Id is not constructed as expected and it is not valid.

In addition, we could observed some web log entries which recorded from search engines and not directly involve with the client user. (Ex: //www.baidu.com/search.) We dropped those web log entries from the initial data set as we are not interested in web page access patters initiated from search engines.

With related to our objectives, we consider only the name of web pages presented in a given web log record. To identify this web page, we manipulate web page name from uri stem attribute and filtered only .aspx pages which will help to extract the real meaning of data.

In initial data exploratory analysis, we identified some web log records with **404.aspx** (The page cannot be found) web page as well. As, these web log records are not attracted to our objective, we dropped those records from the initial data set.

session_Id	date	time	date_time	cs_ip	web_page	time_taken	sc_status	sc_substatus
-1811938738	2014-08-01	00:00:25	2014-08-01 00:00:25.0	202.140.108.99	LocationContacts.aspx	19	200	0
-1442062705	2014-08-01	00:00:28	2014-08-01 00:00:28.0	207.6.118.176	home.aspx	210	302	0
-1738057240	2014-08-01	00:00:40	2014-08-01 00:00:40.0	14.199.63.188	default.aspx	20	200	0
-1442062705	2014-08-01	00:00:48	2014-08-01 00:00:48.0	207.6.118.176	GuestRooms.aspx	327	200	0
-1738057240	2014-08-01	00:00:48	2014-08-01 00:00:48.0	14.199.63.188	default.aspx	20	200	0
1674396148	2014-08-01	00:01:12	2014-08-01 00:01:12.0	61.92.230.63	404.aspx	27	404	0
1561145465	2014-08-01	00:01:30	2014-08-01 00:01:30.0	119.236.43.88	Dining.aspx	35	200	0
1561145465	2014-08-01	00:01:31	2014-08-01 00:01:31.0	119.236.43.88	404.aspx	27	404	0
133623124	2014-08-01	00:01:41	2014-08-01 00:01:41.0	61.92.230.63	LocationContacts.aspx	53	200	0
1674396148	2014-08-01	00:01:44	2014-08-01 00:01:44.0	61.92.230.63	404.aspx	26	404	0

We could observed entry such as "layouts/Layouts/Hotel_ICON_revamp.aspx" which does not interpret the interest of webpage patters. Hence, we ignore all the web page log records include "layouts/Layouts/Hotel_ICON_revamp.aspx"

session_Id	date	time	date_time	web_page	time_taken	sc_status	sc_substatus
-2147474918	2014-10-30	11:36:47	2014-10-30 11:36:...	Hotel_ICON_revamp...	113	200	0
-2147474918	2014-10-30	11:36:47	2014-10-30 11:36:...	Hotel_ICON_revamp...	53	200	0
-2147454191	2014-10-07	06:41:43	2014-10-07 06:41:...	events.aspx	632	200	0
-2147445233	2014-10-18	09:12:40	2014-10-18 09:12:...	offers.aspx	1060	200	0
-2147405036	2014-09-01	05:59:56	2014-09-01 05:59:...	dining.aspx	45	200	0
-2147358177	2014-09-02	23:38:30	2014-09-02 23:38:...	dining.aspx	70	200	0
-2147325756	2014-09-25	19:47:48	2014-09-25 19:47:...	offers.aspx	1848	200	0
-2147325756	2014-09-25	19:50:04	2014-09-25 19:50:...	offers.aspx	3895	200	0
-2147323665	2014-08-11	00:04:29	2014-08-11 00:04:...	rooms.aspx	7188	200	0
-2147313159	2014-09-22	06:30:13	2014-09-22 06:30:...	offers.aspx	1547	200	0
-2147313159	2014-09-22	06:52:18	2014-09-22 06:52:...	offers.aspx	1316	200	0
-2147313159	2014-09-22	06:54:50	2014-09-22 06:54:...	offers.aspx	1364	200	0
-2147313159	2014-09-22	06:57:49	2014-09-22 06:57:...	offers.aspx	1466	200	0
-2147313159	2014-09-22	07:00:22	2014-09-22 07:00:...	offers.aspx	1465	200	0
-2147285169	2014-10-28	16:53:04	2014-10-28 16:53:...	Hotel_ICON_revamp...	1502	200	0
-2147285169	2014-10-28	16:53:05	2014-10-28 16:53:...	Hotel_ICON_revamp...	460	200	0
-2147285169	2014-10-28	20:30:42	2014-10-28 20:30:...	Hotel_ICON_revamp...	1214	200	0
-2147282320	2014-10-28	07:02:14	2014-10-28 07:02:...	Hotel_ICON_revamp...	201	200	0
-2147282320	2014-10-28	07:57:29	2014-10-28 07:57:...	Hotel_ICON_revamp...	224	200	0
-2147206189	2014-08-21	16:11:13	2014-08-21 16:11:...	dining.aspx	100	200	0

Interpretation of Session

To make the constructed session Id standardize, we use hash function to manipulate the session ID based on identified attributes. The sample of generated session Id are as below.

session_Id	time	web_page	time_taken	sc_status	sc_substatus
-1811938738	00:00:25	LocationContacts.aspx	19	200	0
-1442062705	00:00:28	home.aspx	210	302	0
-1738057240	00:00:40	default.aspx	20	200	0
-1442062705	00:00:48	GuestRooms.aspx	327	200	0
-1738057240	00:00:48	default.aspx	20	200	0
1561145465	00:01:30	Dining.aspx	35	200	0
133623124	00:01:41	LocationContacts.aspx	53	200	0
89641610	00:01:47	offers.aspx	129	200	0
944277774	00:02:04	Dining.aspx	25	200	0
89641610	00:02:35	dining.aspx	104	200	0
1003972523	00:03:29	privacy.aspx	219	200	0
1835046587	00:03:29	offers.aspx	250	200	0

Sample dataset for Analysis

After applying feature section and data reduction steps, the sample set of web log data displays as follow for a given session Id.

session_Id	web_page	time_taken	sc_status	sc_substatus	Sequence
-1433039757	offers.aspx	282	200	0	1
-1433039757	offers.aspx	399	200	0	2
-1433039757	offers.aspx	176	302	0	3
-1433039757	offers.aspx	148	200	0	4
-1433039757	offers.aspx	308	200	0	5
-1433039757	above-and-beyond.aspx	88	200	0	6
-1433039757	offers.aspx	347	200	0	7
-1433039757	facilities.aspx	309	200	0	8
-1433039757	rooms.aspx	101	200	0	9
-1433039757	rooms.aspx	224	200	0	10
-1433039757	about-the-hotel.aspx	244	200	0	11

Sample dataset for Analysis of user patterns

It could observed that continuation of same webpage web log data were recorded for given session. As we are interested on web page patterns, we filtered out data set and sample flow is as below.

session_Id	web_page	Sequence
-1433039757	offers.aspx	1
-1433039757	above-and-beyond.aspx	2
-1433039757	offers.aspx	3
-1433039757	facilities.aspx	4
-1433039757	rooms.aspx	5
-1433039757	about-the-hotel.aspx	6

Analysis and results

Frequent pattern Identification – Data Mining

For the month of August 2014, the frequents patters were identified using FPGrowth algorithm as below.

Frequent web-page patterns

Based on the analysis, among users of hotel TULIP the most frequent page visit patterns identified throughout the month is highlighted below. Interpretation of this is like among the all users, number of users who has visited offers page and dinning page in one session is 2456. Similarly, number of users, who have visited facilities page, room's page and offers page in single session is 791.

```
FreqItemset(items=[u'guestrooms.aspx'], freq=1927)
FreqItemset(items=[u'dining.aspx'], freq=11974)
FreqItemset(items=[u'locationcontacts.aspx'], freq=1682)
FreqItemset(items=[u'offers.aspx'], freq=9366)
FreqItemset(items=[u'offers.aspx', u'dining.aspx'], freq=2456)
FreqItemset(items=[u'rooms.aspx'], freq=6369)
FreqItemset(items=[u'rooms.aspx', u'offers.aspx'], freq=2105)
FreqItemset(items=[u'rooms.aspx', u'dining.aspx'], freq=1256)
FreqItemset(items=[u'our-city.aspx'], freq=1057)
FreqItemset(items=[u'aboutus.aspx'], freq=703)
FreqItemset(items=[u'facilities.aspx'], freq=4619)
FreqItemset(items=[u'facilities.aspx', u'rooms.aspx'], freq=1563)
FreqItemset(items=[u'facilities.aspx', u'rooms.aspx', u'offers.aspx'], freq=791)
FreqItemset(items=[u'facilities.aspx', u'rooms.aspx', u'dining.aspx'], freq=747)
FreqItemset(items=[u'facilities.aspx', u'offers.aspx'], freq=1503)
FreqItemset(items=[u'facilities.aspx', u'offers.aspx', u'dining.aspx'], freq=744)
```

Frequent Web Pages

According to frequent pages, dining.aspx and offers.aspx web pages were frequently accessed during whole month by users of hotel Tulip.

According to frequent pages, press.aspx web page was least frequently accessed during whole month by users of hotel Tulip.

```
FreqItemset(items=[u'guestrooms.aspx'], freq=1927)
FreqItemset(items=[u'dining.aspx'], freq=11974)
FreqItemset(items=[u'locationcontacts.aspx'], freq=1682)
FreqItemset(items=[u'offers.aspx'], freq=9366)
FreqItemset(items=[u'offers.aspx', u'dining.aspx'], freq=2456)
FreqItemset(items=[u'rooms.aspx'], freq=6369)
FreqItemset(items=[u'rooms.aspx', u'offers.aspx'], freq=2105)
FreqItemset(items=[u'rooms.aspx', u'dining.aspx'], freq=1256)
FreqItemset(items=[u'our-city.aspx'], freq=1057)
FreqItemset(items=[u'aboutus.aspx'], freq=703)
FreqItemset(items=[u'facilities.aspx'], freq=4619)
FreqItemset(items=[u'facilities.aspx', u'rooms.aspx'], freq=1563)
FreqItemset(items=[u'facilities.aspx', u'rooms.aspx', u'offers.aspx'], freq=791)
FreqItemset(items=[u'facilities.aspx', u'rooms.aspx', u'dining.aspx'], freq=747)
FreqItemset(items=[u'facilities.aspx', u'offers.aspx'], freq=1503)
FreqItemset(items=[u'facilities.aspx', u'offers.aspx', u'dining.aspx'], freq=744)
FreqItemset(items=[u'facilities.aspx', u'dining.aspx'], freq=1524)
FreqItemset(items=[u'about-the-hotel.aspx'], freq=3766)
FreqItemset(items=[u'about-the-hotel.aspx', u'rooms.aspx'], freq=1416)
FreqItemset(items=[u'about-the-hotel.aspx', u'offers.aspx'], freq=1146)
FreqItemset(items=[u'about-the-hotel.aspx', u'facilities.aspx'], freq=1061)
```

Visualization of total visitors by their countries

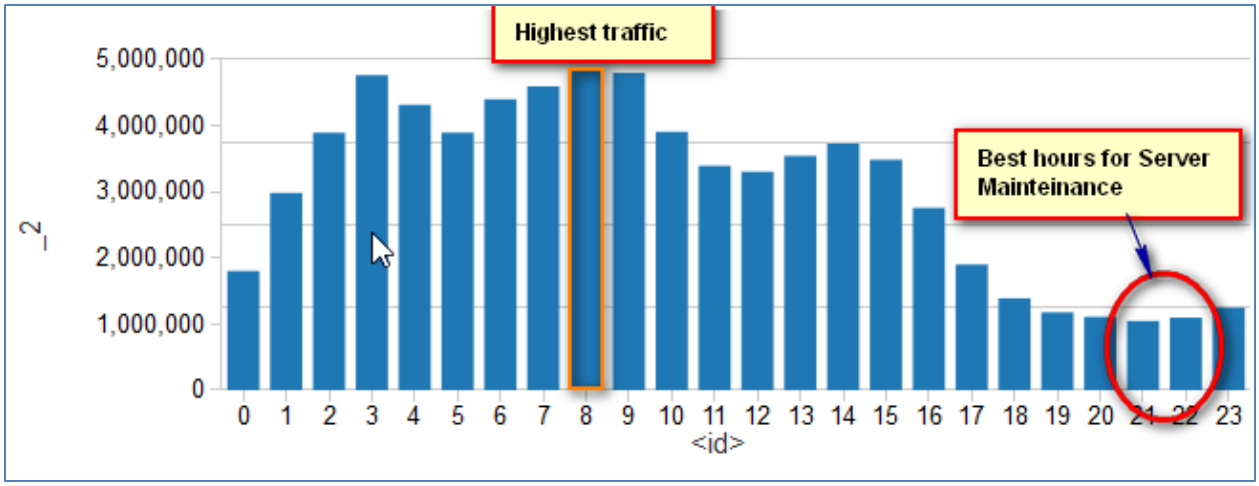
ip	cca2	cca3	cn
105.235.137.22	DZ	DZA	Algeria
105.111.108.88	DZ	DZA	Algeria
41.111.94.255	DZ	DZA	Algeria
185.25.49.181	LT	LTU	Lithuania
203.81.94.89	MM	MMR	Myanmar
122.248.100.148	MM	MMR	Myanmar
122.248.102.196	MM	MMR	Myanmar
122.248.101.221	MM	MMR	Myanmar

Using the “urllib2” and “iso3166” packages , we have mapped cs_ip addresses to their cca2 and cca3 country codes. And then using this data, world map has been derived for better visualization.

TULIP hotel server has got maximum visitors from US and China which are highlighted above.

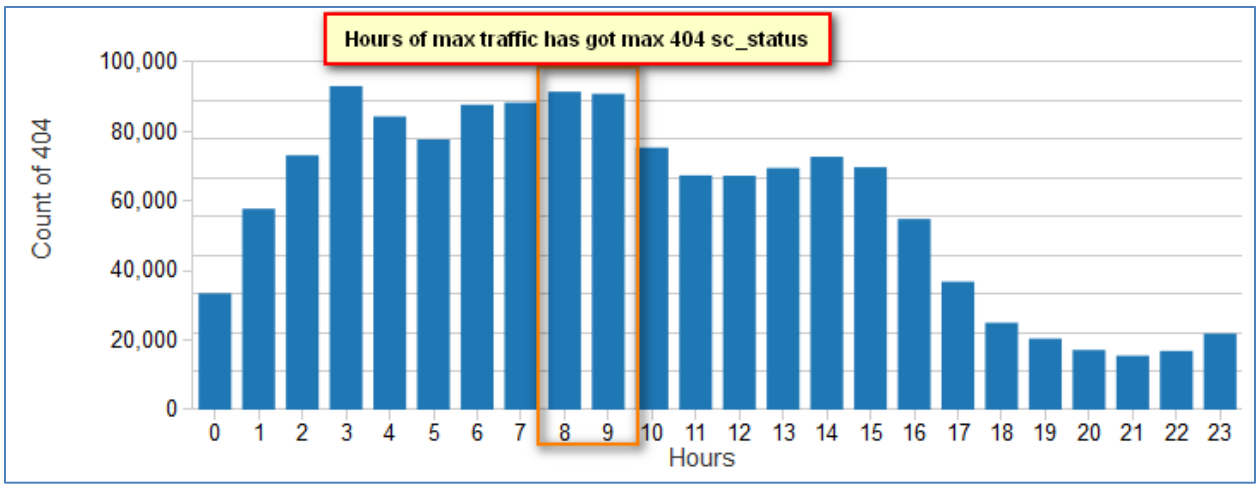
Best Server maintenance hours of the day

By analyzing the server access records, it is identified that the 9:00 pm to 10:00 pm is the best time for any kind of server maintenance work as server has got minimum traffic in this duration. While, from the same it is justified to see that during 8:00 am to 9:00 am , server has got highest traffic during this time



Maximum 404 during busiest Hours

After analyzing the 404 server status code and it's frequency during the whole day , it is identified that Server was unable to satisfy all the requests during the time of maximum requests(8:00-9:00 pm) which should be taken into account to improve the server resources.



Conclusive Business Insights from analysis

Conclusively, after analyzing the server log access data for Hotel TULIP , we have interesting and useful outcomes that we can suggest to IT Division of Hotel TULIP.

- Webpages for offers, dining, facilities and rooms are most frequently accessed pages. So these webpages can be enhanced to improve user experience. This can also be useful for the social media and marketing team.
- Countries from where Hotel TULIP is getting very less customers can be taken into consideration for improving their marketing campaign into those countries.

- Best suitable time to do server maintenance can be taken into consideration for better customer satisfaction.
- It's also recognized that during busy hours, server is not capable of handling all the requests successfully so the server resource allocation can be improved specially for maximum traffic hours.

References

CU, O & Bhargavi, P 2013, 'Analysis of Web Server log by web usage mining for extracting users patterns', *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)*, vol. 1, no. 3, pp. 123-36.