CS532-01: Final Project Report

Project Title: Log File Analysis

Team Member(s): Anmol Pal – B00976430

/* Your project will be graded based on the significance of the project and the success of the demo and the clarity of your report. Your final report excluding the references and source code should not exceed 3 pages. */

/* Please organize your report well for clarity, and always check spelling and grammar. An excellent style manual for science writers is [2]. */

I. Problem

Log File Analysis: Using Apache Webserver logs to understand the behavior of an application and trends of the customer base. Almhuette-raith[5] is a small German Website. The business provides services of renting lodges to customers and their website allows interaction with customers. The log analyzer takes Almhuette-raith's webserver logs[6] to perform the below two analysis tasks.

1. Behavior of application: The trends of traffic received on the application for over a year.
2. Understanding customer base: Obtaining insights on customers' geographic location and identifying trends if any.

II. Software Design and Implementation

A. Software Design and NoSQL-Databse and Tools Used:

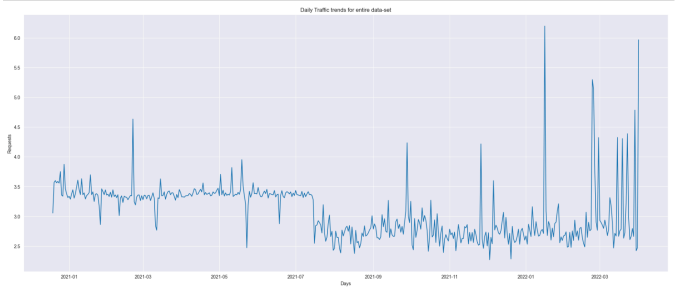
The application is divided into 3 parts once logs have been extracted.

1. Data Wrangling[1] – Webserver logs in raw text format are transformed to Apache Spark [2] Data Frame format, using Pandas APIs on spark. Eliminating unnecessary columns concerning the scope of the analysis tasks in this project.
2. Creating more columns based on information derived from the dataset: Transforming the Host (Customer) IP address to the geographic location of origin using batch processing of the Host column to a CSV file, using ip-api[7] for this transformation used. Used Python's Pandas[3] library to perform the task.
3. Analytics: The newly modeled data frames are then used to perform analytics. Visualization of customer traffic using matplotlib[8] , seaborn[9], plotly[10] to identify the weekly , monthly and geographic trends.

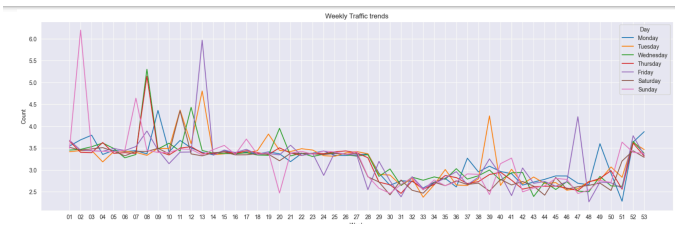
B. All tasks mentioned above implemented by Anmol Pal

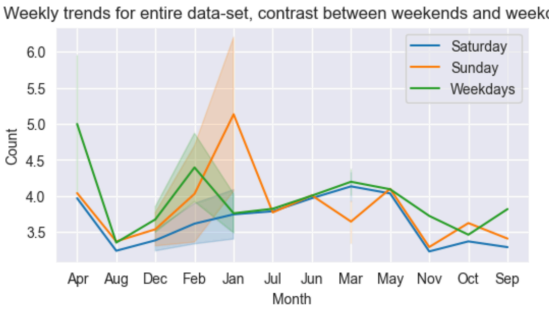
III. Project Outcome

1. A clean data set to work with.
2. Transformation of information available in Host IP to Geo Location, Timestamps to several Date Time instances to work by weeks and months upon the data-set.
3. Analysis and deducing business insights upon visualizing the data.
4. Business Insight: We observe requests reducing drastically in mid-2021 – Wave 2 during the pandemic and spikes back in 2022 when tourism was active at its peak.

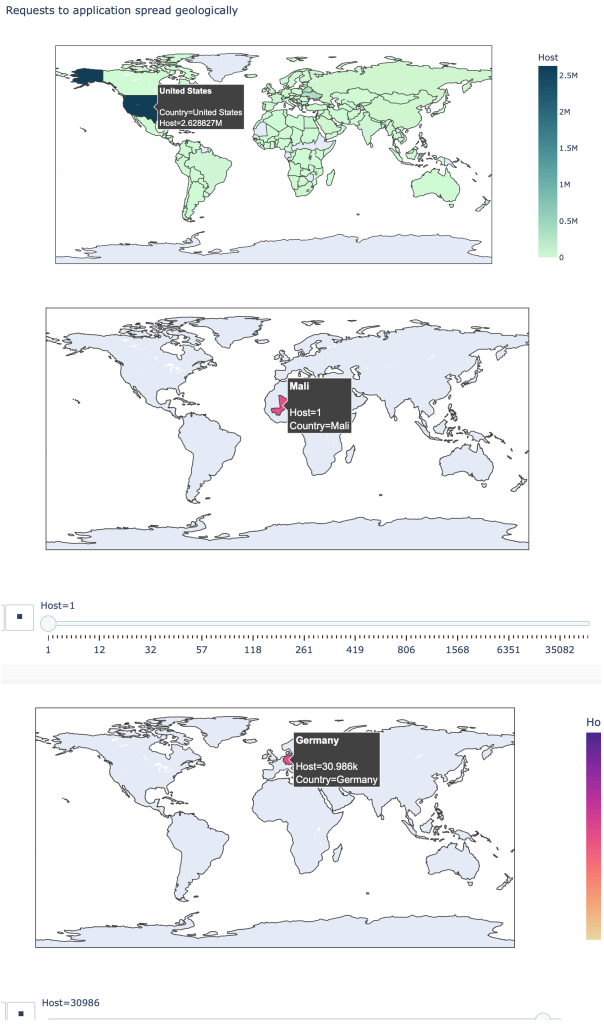


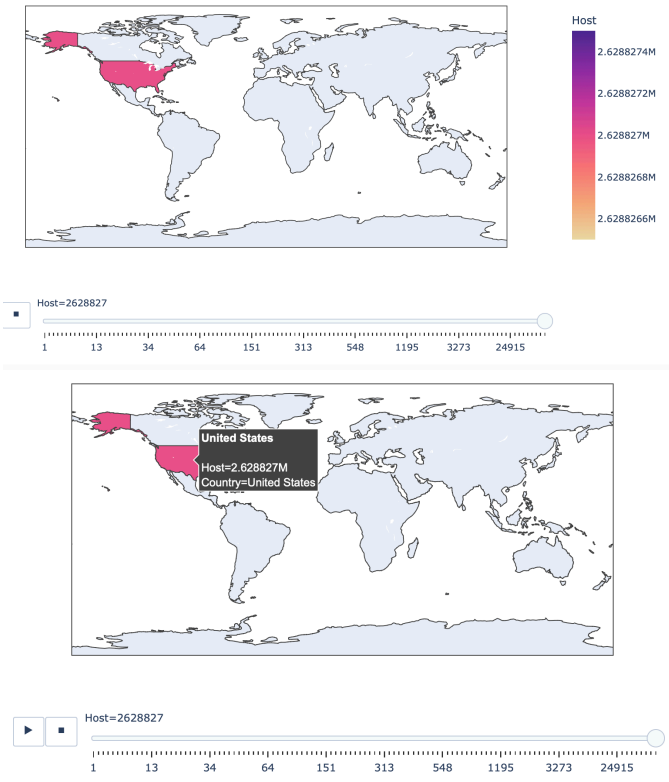
5. Weekly Trends: Frequency of requests by weekdays throughout the year. We can see Saturdays and Sundays have significantly higher traffic as compared to the rest of the week.





6. Upon occasions of spikes in the loads on the website, one might consider scaling up the infrastructure for the website to handle the load well as the requests go up to millions. Furthermore, important business insight is Almhuette-Raith can opt for different strategies to expand their business during the peak days to maximize profit.
7. Geographic Analysis of customer base:
We can identify a lower customer base from locations like Mali, Zambia, Iceland, etc., whereas maximum traffic from the United States, Russia, South Africa, etc.





8. Most useful business insight that can be drawn from the location of the customer base is that lodges being offered can be made suited more for such customers based on their culture and needs which will, in turn, will give the brand a great reputation and help the business thrive.

REFERENCES

[1] Data Wrangling : https://opensource.com/article/19/5/log-data-apache-spark

[2] Python Spark Api : https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_df.html

[3] Pandas on spark: https://spark.apache.org/docs/latest/api/python/getting_started/quickstart_ps.html

[4] Geo-Maps Chrolopleth: https://plotly.com/python/choropleth-maps/

[5] Almhuette-raith : http://www.almhuette-raith.at/index.php?option=com_content&view=article&id=49&Itemid=55

[6] Almhuette-raith's webserver logs: http://www.almhuette-raith.at/apache-log/access.log

[7] http://ip-api.com/batch

[8] matplotlib : https://matplotlib.org/cheatsheets/_images/cheatsheets-1.png

[9] Seaborn: http://seaborn.pydata.org/

[10] Plotly: http://plotly.com/

[11]