Data Visualization on GitHub repository parameters using Elastic search and Kibana

Mohan Kumar J
Manipal Academy of Higher Education, Manipal
mohan.js@manipal.edu

Shishir Dubey
Manipal Academy of Higher Education, Manipal
er.shishirdubey@gmail.com

Balaji B
Manipal Academy of Higher Education, Manipal
balaji.b@manipal.edu

Dinesh Rao
Manipal Academy of Higher Education, Manipal
dinesh.rao@manipal.edu

Deepak Rao
Manipal Academy of Higher Education, Manipal
deepak.rao@manipal.edu

Abstract: Data Visualization makes data mean more through storytelling. Any data will have an inside story when it is addressed with relevant query. In this paper, GitHub data is collected, cleansed and visualized for its repository parameters on various questions. GitHub being the platform for many developers is not only used for version controlling their project but also for sharing it. Through this technique the inside story and the use of GitHub is seen.

Keywords: GitHub; Data Visualization; Big Data.

I. INTRODUCTION

Data visualization is a promient and efficient tool for story telling. John Snow's cholera map [1][4] data visualization is an very good example to understand the power of Data Visualization.Another example of data visualization is the Parisian civil engineer Charles Minard's [2] [4] cartographic depiction of numerical data on a map of Napoleon's disastrous losses suffered during the Russian campaign of 1812. In. The Crimean War 1855, Britain was fighting a battle with both Russia and a disease. As a nurse, it was a great effort putforth by Florence Nightingale, to convince an army to invest in hospitals and healthcare instead of guns and ammunition. Nightingale [3] [4] detailed the story with data by showing the staggering amount of deaths due to the preventable disease. After this visualization, sanitation became a major priority for the British Army. Many more examples in literature endorse that visualization has helped to reveal the data in a better way than simple or complex representation in numbers. In present,Big data and Internet of things are the technologies which make data visualization as an important tool to be used. These tecnologies have enabled for various research problems. Huge set of data is available over the Internet servers. These data are scrapped, cleaned and analysed. The analysed data is represented in the form of visualization rather numbers in order to get better inference. With GitHub many parameters can be analysed at an user perspective. In order to access these data, GitHub provides API, where developers use it to collect the required information.

II. LITERATURE REVIEW

Found in 2008, GitHub in recent times has been the research point for many researchers. Initially it started with 2000 users, and by now, it has surpassed 25 million users. [5] has explained the data collection using the REST API. In [6], GitHub data collection and mining was done to understand user performance and project achievement factors. A detailed analysis is done on "pull" request using Git in [7], to know whether the developers feel comfortable in using it. A survey is done with research question and visualized. The effects of pull requests are early studied and modelled in [8]. The social network based user projects are analyzed in [9]. in which the potential health of a project is concluded to be 40-50%. In [10] the factors like programming languages and the application domains are analyzed. Around 2500 public repositories were used for the purpose with high stars in GitHub. It concludes, compared to the individual owned repositories, the organization owned repositories are more popular. In [11], the author tries to predict the repositories in which developers will join. Machine learning classification techniques is applied to predict it. A novel approach is proposed on finding similar repositories in GitHub in [12]. The work is based on two data sources, GitHub and the readme files. In the study, it reveal that RepoPal [12], an integration system build with three scores namely, readme-based, stargazers based and time-line based. [13] study is done on how software development practices are influenced by social coding with platform like GitHub. The information is extracted from GitHub archive for 459 projects, and the contents of a contribution file and their relationships are analyzed for the contents and acquired contributors [14].

III. Methodology

The methodology adopted is same as any other data visualization technique which involves Data Collection, Data Exploration, Data Preprocessing and Visualization. The metadata of GitHub uses truly random samples of around 500 users. The collected data is locally cached using the NOSQL database tool to speed up analytical tasks. The metadata collected is explored using exploratory data analysis; and analyzed to get the interesting insights of the GitHub users. A few research questions(RQ) were framed to work with the data. For the RQ's the results were visualized using different visualization techniques. The RQ's are:

1. Most Frequent Domains by Description
2. Highlighting the popularity of Languages per region by Repositories Count
3. Popular Companies by Stargazers Count and Top 5 Languages
4. Infering Technically Developed, Developing and Underdeveloped cities or countries
5. Popularity of companies by Repositories
6. Busy Weekdays According to the GitHub

Data was collected using a custom python application. The application includes three separate modules for the collection of data. The first module involves the collection of random users. The second module implies the collection of repositories for respective users. The third module contains the collection of commit logs for respective repositories. Elasticsearch [16] and Kibana[17] is used for visualization [15].

IV. Results

RQ1: Most dominant Domains on GitHub by Repository Description

Figure 1. shows the distribution of the domains which are highly popular according to the description of the repository. The top-9 domains are Plugin (with 864 repositories), web (with 815 repositories), Frameworks (with 777 repositories), API (with 693 repositories), Data (with 626 repositories), Applications (with 603 repositories), Mobile Applications (with 519 repositories), System Files (with 458 repositories), and server (with 435 repositories).

RQ2: The popularity of languages per Region by Repositories Count.

At worldwide, there are several developers and users who utilize different languages to perform different tasks as per their compatibility, comfortability and stability which may vary from region to region; and to know about the variance of languages per region, this analysis is been made based on repository counts. Here the highest repository count of every region is considered with the language used within those repositories which shows the most dominant language within that region.

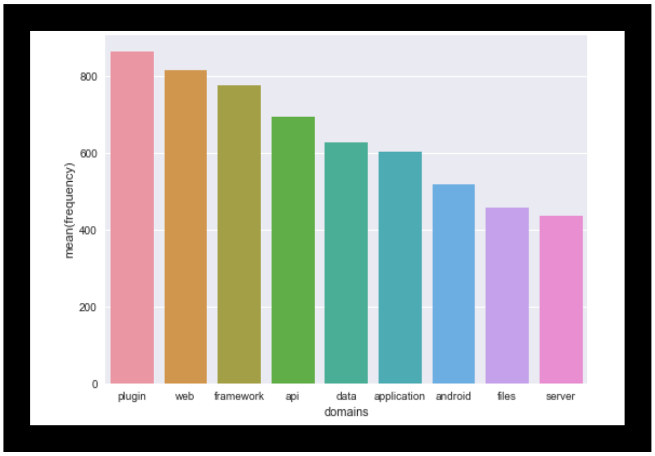

Fig.1. Most Frequent Domains by Description

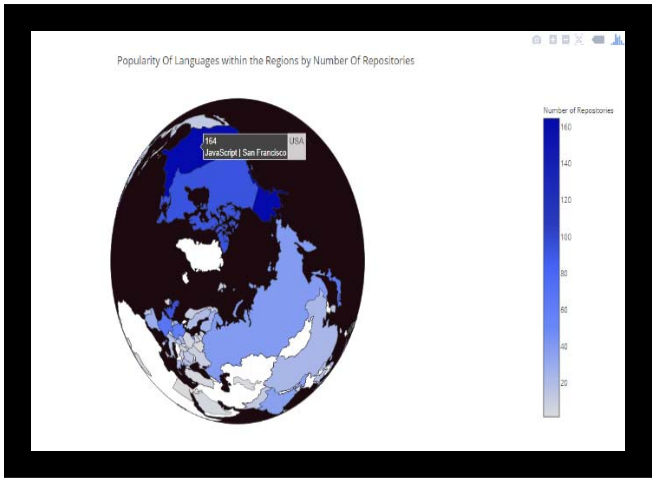
Popularity Of Languages within the Regions by Number Of Repositories
164
JavaScript | San Francisco
USA
Number of Repositories
160
140
120
100
80
60
40
20

Fig. 2. Highlighting the popularity of Languages per region by Repositories Count

As here in the Figure 2, color bar represents the variance of counts of repository on a language. Higher the repositories count darker the color within the map and lower repositories count will represent light color within the map. As shown in the Figure 2., if users wanted to see the popular languages within the region, they can simply hover on the map which will show the highest repositories count with the name of the city and country's ISO code. As the highest repository count of 164 repositories is found for the region USA it is darker in the Map, and JavaScript is been highly used within the repositories in USA which infers that JavaScript is the most popular language within USA. For the visualization implementation, python nvd3 library [18] is used.

RQ3: Popularity of Companies by Stargazers Count and Top 5 Languages

Many companies work upon a same domain but follow different tools and techniques within the companies for either gaining the performance or customer's attention. The growth factor for a company, can be either by keeping a good track of their work or by innovating new technologies or maintain decent quality of products. These kinds of companies gain followers and stargazers who used to follow them for their latest updates and give stars according to the quality of companies, thereby increasing its popularity. The stars denote the popularity of the companies. More the number of stars more the popularity. Here the analysis is being made upon GitHub repositories data to know the popular companies and top 5 languages through stargazers count. As here in Figure3, the inner circle represents the top 5 languages which are Ruby, Python, PHP, Perl, JavaScript and the outer circle represents the top companies like beyond, rent.com, google, phone master using those languages with respective stargazers count. Here, both languages and companies are grouped by Stargazers Count. The color bar shows the respective color related to companies or languages. As it is an interactive plot, the users can hover the mouse on the fields to filter the language and companies.

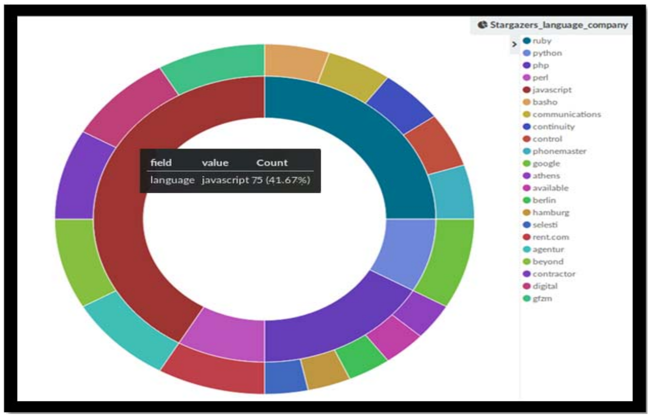

Fig.3. Popular Companies by Stargazers Count and Top 5 Languages

RQ4: Inferring Technically Developed, Developing and Underdeveloped cities or countries

Though there are many insights of GitHub being discovered in the previous studies, there is no discussion been made for finding the technically developed countries, technically developing countries and underdeveloped countries in accordance with the GitHub Data. It is an interesting analysis for getting to know which country stands strong technically and done based on stargazers count. In Figure 4, the red circles are the top 9 countries or cities which have highest stargazers count, the green circle shows the technically developing countries or cities with the average stargazer counts (10-991) and the values which are hardly visible on the map with the blue color represents the rest of the remaining countries or cities with very less stargazer counts lesser than 10.

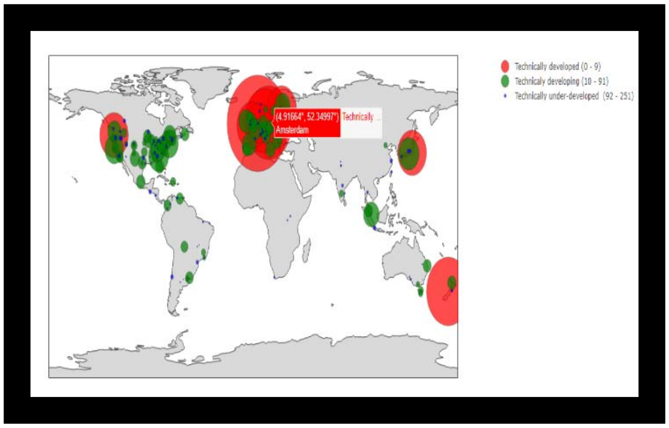
Technically developed (0 - 9)
Technicaly developing (10 - 91)
Technically under-developed (92 - 251)
(4.91664°, 52.34997°) Technically ...
Amsterdam

Fig.4.Technically Developed, developing and under-developed countries

In the map shown above, Wales is the one with biggest red circle - indicating the highest watchers counts and portrays itself as a developed city within all European countries and cities. The green circles in the European counties shows that Europe has highly technical developing countries than other regions.

It is also an Interactive plot which allows the user to filter 3 of the categories - developed, developing, and underdeveloped countries, with the help of the color bar shown near the plot.

RQ5: Popularity of companies by Repositories

Figure 5. shows the distribution of top 20 popular companies with the highest repository counts along with the languages used by those companies in a stack. As shown in the Figure 5, every company almost uses C and JavaScript for their repositories. Other languages used within the repositories are PHP, Java, Ruby and Python.

Coffee script and action script is rarely used by some companies for the development. We can infer that the popular companies are still using C and adapting slowly to new languages.

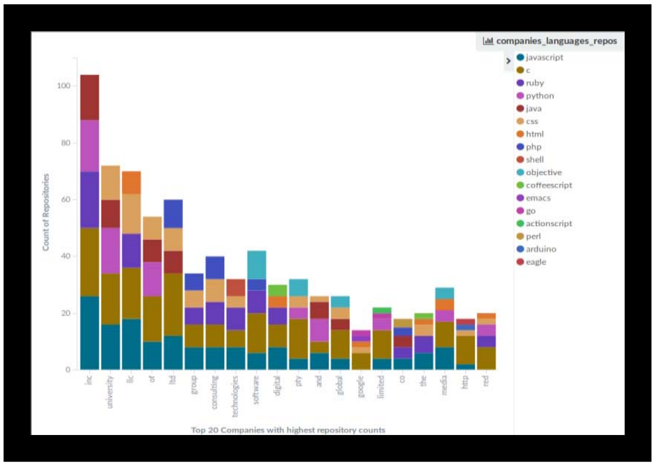

Fig.*5.* Top 20 Popular companies by Repository

RQ6: Busy Weekdays According to the GitHub

Here analysis is made upon the field “pushed at” which contains the Date and time of the pushes. Some transformations have been made on the field to get the weekdays of the pushes made, which gave the idea of finding the busy weekdays in GitHub.

As in Figure 6, it is clearly visible that “Tuesday” is the busiest day where users used to push more changes or files. The second busiest day is Wednesday followed by Monday.

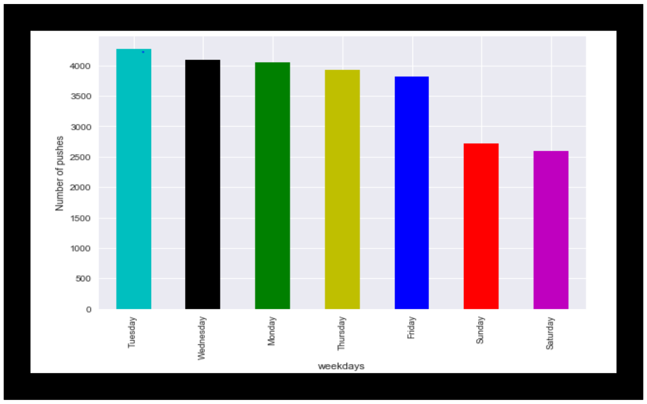

Fig. 6. Busy Weekdays according to the GitHub Data.

From the fourth day of the week users have less work to do which is Thursday and Friday. Usually on Saturday and Sunday users prefer to take rest from the work and rarely used to push changes or files.

V. CONCLUSION

GitHub has become one of the familiar code hosting server application for individualsas well as organizations. From the data collected, few RQ’s were handled and visualized. . A predictive model built on the GitHub collected data can be the scope for the future research.

ACKNOWLEDGMENT

We acknowledge Dr. Deviprasad for his constant support and inputs provided, in completion of the work.

REFERENCES

[1] https://en.wikipedia.org/wiki/John_Snow

[2] https://en.wikipedia.org/wiki/Charles_Joseph_Minard

[3] https://en.wikipedia.org/wiki/Florence_Nightingale

[4] http://www.tableau.com/sites/default/files/whitepapers/the_5_most_influential_data_visualizations_of_all_time.pdf?ref=lp&signin=1b23e5076212b10eab55674fb4d4a226

[5] Georgios Gousios, MSR '13. The GHTorent dataset and tool suite(2013)

[6] Chatziasimidis, Fragkiskos, and Ioannis Stamelos. "Data collection and analysis of GitHub repositories and users." In Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on, pp. 1-6. IEEE, 2015.

[7] Saito, Yusuke, Kenji Fujiwara, Hiroshi Igaki, Norihiro Yoshida, and Hajimu Iida. "How do GitHub users feel with pull-based development?." In Empirical Software Engineering in Practice (IWESEP), 2016 7th International Workshop on, pp. 7-11. IEEE, 2016.

[8] Liu, Jing, Jiahao Li, and Lulu He. "A comparative study of the effects of pull request on github projects." In Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual, vol. 1, pp. 313-322. IEEE, 2016.

[9] Leibzon, William. "Social network of software development at GitHub." In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, pp. 1374-1376. IEEE, 2016.

[10] Borges, Hudson, Andre Hora, and Marco Tulio Valente. "Understanding the factors that impact the popularity of GitHub repositories." In Software Maintenance and Evolution (ICSME), 2016 IEEE International Conference on, pp. 334-344. IEEE, 2016.

[11] Nielek, Radoslaw, Oskar Jarczyk, Kamil Pawlak, Leszek Bukowski, Roman Bartusiak, and Adam Wierzbicki. "Choose a Job You Love: Predicting Choices of GitHub Developers." In Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on, pp. 200-207. IEEE, 2016.

[12] Zhang, Yun, David Lo, Pavneet Singh Kochhar, Xin Xia, Quanlai Li, and Jianling Sun. "Detecting similar repositories on GitHub." In Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on, pp. 13-23. IEEE, 2017.

[13] Cosentino, Valerio, Javier L. Cánovas Izquierdo, and Jordi Cabot. "A Systematic Mapping Study of Software Development with GitHub." IEEE Access 5 (2017): 7173-7192.

[14] Kobayakawa, Naoki, and Kenichi Yoshida. "How GitHub Contributing, md Contributes to Contributors." In Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, vol. 1, pp. 694-696. IEEE, 2017.

[15] Bajer, Marcin. "Building an IoT Data Hub with Elasticsearch, Logstash and Kibana." In Future Internet of Things and Cloud Workshops

(FiCloudW), 2017 5th International Conference on, pp. 63-68. IEEE, 2017.

[16] https://www.elastic.co/products/elasticsearch

[17] https://www.elastic.co/products/kibana

[18] https://pypi.org/project/python-nvd3/