Analyzing Open Source GitHub Repositories Towards Technology Acceptance Model

(Draft 2)

Dhruvil Gandhi
Seidenberg School of Computer Science and Information Systems
Pace University, New York, USA
Email: dgandhi@pace.edu

***Abstract*—Open Source is a gift and effort of the technology community, over years creating and supporting platforms, projects etc. This paper analyzes data from GitHub public repositories available on Google's BigQuery and observes relations, trends and anomalies. To set a baseline, public repositories for 22 different languages is analyzed, coorelations, anomalies and trends are studied.**

***Index Terms*—Technology Acceptance Model, Data Analytics, Time Series Anomalies, Data Exploration**

I. Introduction

Open Source allows community from novice to advanced, everyone to collaborate and contribute to projects, platforms, programming languages and framework. Overtime the advancement of languages and platforms; languages from Assembly Language to C to Python to TypeScript and Kotlin, for platforms - from Batch systems to multiprocessing and modern distributed systems. With evolution of services and platforms, IaaS, Saas and PaaS, lead to growth of open-source community and projects over time.

Various studies have been conducted overtime to study the imapct and growth of these. This study [1] studies the influence of software by studying GitHub. Evaluation [2] and [3] analysis observes and introduces Technology Acceptance Model, and develops hypothesis for it. Another such study [4] was conducted to study UTAUT (Unified Theory of Acceptance and Use of Technology) [4] where a study was done on 78 undergraduate students for acceptance and impact of GitHub on their education. Various studies have been performed on commit messages, coding style, pull requests, coding times and other factors available for the dataset from GitHub [5]

The trends of open source repositories on GitHub were studied and trends were established and anomalies were studied using ELK stack and BigQuery. Initially, coorelation of trends and data with StackOverflow, a technology Q/A platform was proposed, which however is moved as next steps and future action for this study.

Methodology section talks about the technology used, dataset and analysis proposed. System and Experiment sections elaborates the system setup used for the study and presents the steps, algorithms and data analytics methods used in the study. Results, dataset snippet, small data frames, and anomalies are discussed in the observation and result section. The paper ends with preliminary conclusion and states the next steps and opportunities for analysis.

II. Methodology

A public dataset of nearly 3.8 million GitHub open source repositories [6] is available on Google Cloud Platform's BigQuery service. This dataset was used in the study. BigQuery is Google's columnar storage service for high performance and high throughput needs.

To store and analyze data, ElasticSearch and Kibana were used. ElasticSearch is based on Lucene which is a full-text search based search engine. ElasticSearch has advatange of indexing json documents without reindexing, only adding or updating the new or updated documents. To analyze and visualize, Kibana was used, Kibana is a web based UI for ElasticSearch, which provides analysis and other services for ElasticSearch.

To perfrom the study, data about repositories for following languages was loaded in ElasticSearch from BigQuery, using Python. The repositories for which the repositories were loaded are highlighted in table I. The fields that were loaded from the BigQuery dataset are as in table II.

Language Name	Total Repositories
C	5,451,310
Python	878,748
Go	734,652
C++	689,476
Java	626,740
Dockerfile	264,042
Objective-C	144,206
Haskell	52,430
Clojure	43,686
Rust	34,519
R	29,813
Jupyter Notebook	26,891
Erlang	25,694
TypeScript	14,372
Julia	13,529
Kotlin	12,915
Swift	9,379
Objective-C++	60
Vue	38
SQL	20

TABLE I
LANGUAGES AND REPOSITORIES

Field Name	Data Type
Repository Name	String
Language Name	String
License Name	String
Timestamp of creation	Epoch Time in seconds
Year of creation	Integer
Month of creation	Integer

TABLE II
SCHEMA OF DATA USED

To analyze, wordcloud, average repositories over time for all languages, different visualizations, timeseries analysis and timeseries anomaly detection and forecast were carried out. The result is described in section V and additional materials supplied with the paper.

One limitation while conducting the study is limit of availability of data on percentage of language for a particular repository, for example a simple webpage might have HTML and CSS, but it does not have data on what percent of each language is present in a particular repository.

III. System and Experiment

IV. Observation and Result

7 images and graphics outlier to conference/event etc

V. Future Work

more parameters and data from stack and mention of repositories in stack

VI. Conclusion

pattern, timeseries, relation

Acknowledgment

The authors would like to thank...

References

[1] A. S. Badashian and E. Stroulia, “Measuring user influence in github: The million follower fallacy,” in *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering*, ser. CSI-SE ’16. New York, NY, USA: ACM, 2016, pp. 15–21. [Online]. Available: http://doi.acm.org/10.1145/2897659.2897663

[2] B. Szajna, “Empirical evaluation of the revised technology acceptance model,” *Management Science*, vol. 42, no. 1, pp. 85–92, 1996. [Online]. Available: https://doi.org/10.1287/mnsc.42.1.85

[3] W. R. King and J. He, “A meta-analysis of the technology acceptance model,” *Information and Management*, vol. 43, no. 6, pp. 740 – 755, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378720606000528

[4] A. Čižmešija and Z. Stapić, “Github as backbone in software engineering course: Technology acceptance analysis,” in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2019, pp. 742–746.

[5] G. Gousios, “The ghtorrent dataset and tool suite,” in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR ’13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 233–236. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487085.2487132

[6] “Github activity data.” [Online]. Available: https://console.cloud.google.com/marketplace/details/github/github-repos?filter=solution-type:dataset&id=46ee22ab-2ca4-4750-81a7-3ee0f0150dcb