Analyzing Open Source GitHub Repositories Towards Technology Acceptance Model

Dhruvil Gandhi

Presented for CS 816

https://github.com/dhruv857/tam816

Agenda

1. Introduction
2. Proposed Experiment
3. Experiment Set-up
4. Results
5. Conclusion and Future work
6. References

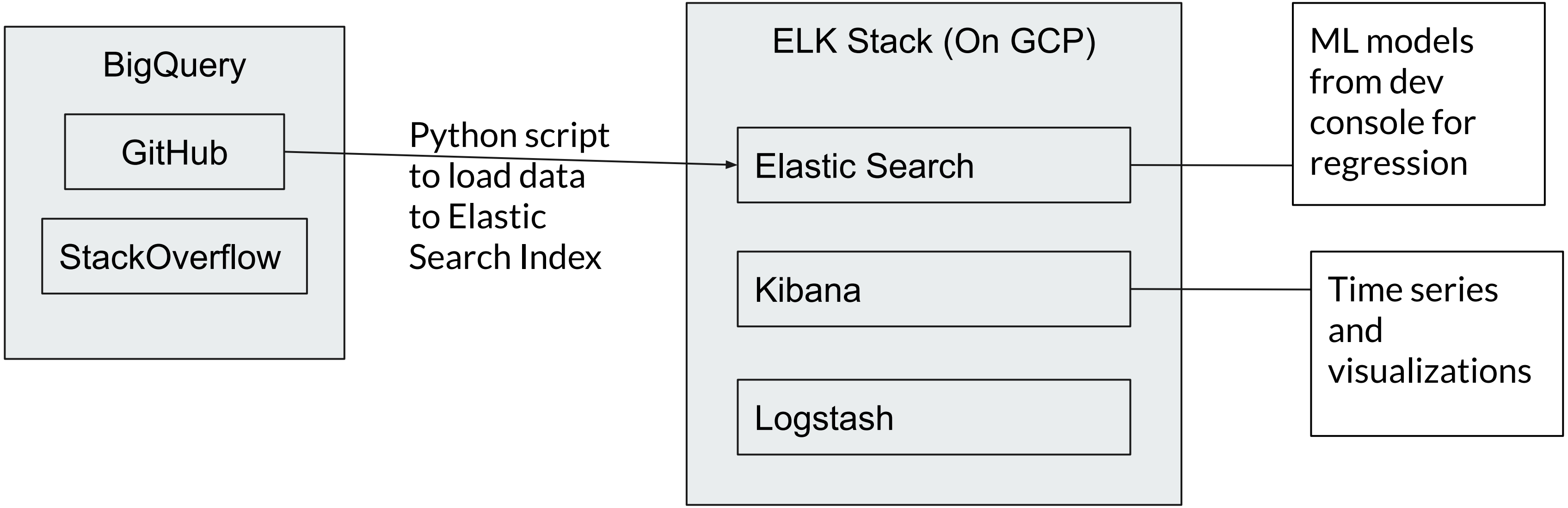
1. Introduction

- Multiple studies have been conducted to study trends and predict acceptance and adaptability of different programming languages
- Different parameters such as commit messages in GitHub, comments in code, questions and answers in Stack Overflow, their textual analysis have been done.
- One such study is done at University of Victoria, which explores prediction programming language.
- Several analyses have been done on the publicly available datasets of 2.8 million open source repositories on GitHub and Q/As of Stack overflow
- We perform regression analysis and time-series analysis for repositories of 20 programming languages using repos, language name and time stamp, running count of repositories.

2. Experiment

- First, I gather all the repositories, languages and creation date from publicly available GitHub open source dataset on Google BigQuery.
- All the data into a single index (document collection) in ElasticSearch
- ElasticSearch’s time-series analysis was used to find trends and anomalies.
- Match anomalies to events and announcements
- Visualize results.

3. Experiment Setup


BigQuery
GitHub
StackOverflow
Python script to load data to Elastic Search Index
ELK Stack (On GCP)
Elastic Search
Kibana
Logstash
ML models from dev console for regression
Time series and visualizations

4. Experiment Links and Results (live-demo)

- http://35.188.72.224/
- Dev Console - Mappings
- TimeSeries
- https://github.com/dhruv857/tam816/blob/master/GitHub%20Language%20Repo%20Analysis.pdf

4. Conclusion

- Trends were observed
- Anomalies were linked with significant event or announcement for a particular programming language.
- Different trends were observed and visualizations were developed for the analyzed data.

4. Future Work

- StackOverflow data for language questions and answers, its time and sentiment analysis.
- GitHub commit messages, time, releases, pull requests, code comments and forking analysis.
- Correlating both datasets with mentions of repositories, issues and links.
- Getting or creating a dataset for significant events for a subset of programming language.

Thank you