GitHub Analytics - Initial Document

Harish Thuwal, Shadab Zafar

October 2017

Dataset

We're using the dataset from a code hosting website called GitHub. The dataset is provided by the GHTorrent project which montiors the GitHub event stream and stores the data. We downloaded the latest dataset, which is a 60 GB compressed tarball that contains 20 CSV files. When uncompressed, the total size of the dataset is 222 GB of plain text. Each file corresponds to table of a MySQL database, the schema of which is available here. Some details of the dataset are given in Table 1.

	File	Size	Data
1	users.csv	1.6 GB	GitHub Users
2	organization_members.csv	16.1 MB	Users that are members of an organization
3	followers.csv	588.7 MB	Users that follow another user
4	watchers.csv	2.9 GB	Users that watch a project
5	projects.csv	11.1 GB	GitHub Projects
6	project_commits.csv	89.8 GB	Commits of projects
7	project_languages.cs	4.3 GB	Programming languages used in projects
8	project_members.csv	491.8 MB	Users that are contributors to projects
9	repo_labels.csv	6.6 GB	Labels used in a project
10	commits.csv	68.8 GB	Commits on projects
11	commit_comments.csv	731.6 MB	Comments made on commits
12	commit_parents.csv	13.7 GB	Parent(s) of commits
13	issues.csv	3.0 GB	GitHub Issues made on projects
14	issue_comments.csv	4.1 GB	Comments made on issues
15	issue_events.csv	4.9 GB	Actions taken on issues (closing etc.)
16	issue_labels.csv	262.6 MB	Labels assigned to issues
17	pull_requests.csv	1.2 GB	GitHub Pull-requests made on projects
18	pull_request_comments.csv	2.7 GB	Comments made on pull-requests
19	pull_request_commits.csv	2.3 GB	Commits made on pull-requests
20	pull_request_history.csv	3.5 GB	Actions taken on pull-requests (merging etc.)

Table 1: GHTorrent Dataset Contents (mysql-2017-09-01.tar.gz)

Analysis

We plan to do the following analyses on the data

1. Correlation between time of day and activities?

   Do users commit more often at some particular time?

   This can be done for various kinds of GitHub activities:

   - Commits
   - Comments
   - Opening of Issues
   - Creation of Pull Requests

2. Lifespan of a project

   - After what time period activities in a github project becomes stagnant or comes below a certain threshold.
   - How, on an average, the frequency of commits varies for a project during its lifespan.
   - Do project based on certain set of languages tend to have longer lifespan than others?

3. Active User Acquisition Rate

   How has the number of active users changed over time? This will also need to consider users that later deleted their accounts or have made no contributions on the site.

4. Developer Countries

   Which countries do developers reside in and how has that data varied over time?

5. Forks

   Is there a correlation between the number of times a project is forked and the languages used in that project?

6. Most Active Users

   Who were the most active users for a range of time, what language they wrote code in. Activities can again be measured with commits, issues etc.
   Other parameters can be the number of followers a user has or the number of organizations he is a member of.

7. How much of a project is community driven?

   What percentage of commits in a project were part of a pull request? or were authored by users who are not the creator of the project.

   After a project's creation, how much time does it take for the first community participation to occur?