**Github Data Analysis**

**Github API** -> <https://developer.github.com/v3/activity/events/>

**Python SDK** <https://github.com/github3py/github3py>

**Github Archive ->** <https://www.githubarchive.org/>

**Google Big Query** **->** <https://cloud.google.com/bigquery/public-data/github>

* **Repos Contents Dataset** -> <https://bigquery.cloud.google.com/dataset/bigquery-public-data:github_repos>
* **GitHub archive dataset** -> <https://bigquery.cloud.google.com/dataset/githubarchive:github>
  + <https://bigquery.cloud.google.com/table/githubarchive:github.timeline>
* **Queries** -> <https://gist.github.com/alysonla/e14c01ec7a0d2823e7317f7b58b22926>

**GH Torrent** **->** <http://ghtorrent.org/gcloud.html>

Tips

* Don’t analyze the main [bigquery-public-data:github\_repos.contents] table — at 1.5 TB, it will *instantly* consume your monthly free terabyte. Use instead the official [bigquery-public-data:github\_repos.sample\_contents] extract (~23 GB), or one of the full language tables I left at [[fh-bigquery:github\_extracts.contents\_\*](https://bigquery.cloud.google.com/dataset/fh-bigquery:github_extracts)].
* How about doing a JOIN between this new dataset and the GitHub Archive to find the most starred files and their patterns? Sample code soon, but see how I played with [GitHub stars and Hacker News](https://www.reddit.com/r/bigquery/comments/3shl0o/qotd_what_else_did_they_star_get_to_know_your/) previously.
* Getting author and committer timezones. We’ll be able to perform some regional analysis here.

In depth Analysis

We ran [a list of queries](https://gist.github.com/alysonla/e14c01ec7a0d2823e7317f7b58b22926) on the datasets above to create the open source section of our [Octoverse](https://octoverse.github.com/) report, but anyone can run an analysis. Here are the results of some of the queries run so far.

* “This should never have happened” has appeared in code comments more than a million times (hear this data point for yourself in [this Changelog episode](https://changelog.com/podcast/209))
* Where does open source happen? [GitHub top countries](https://medium.com/@hoffa/github-top-countries-201608-13f642493773#.beemirtdc) shares which countries have the most open source developers per capita
* How reliable is GitHub? Felipe runs a query to find out in [GitHub reliability with BigQuery](https://medium.com/@hoffa/sla-slo-explored-with-github-and-bigquery-e6a135919a8e#.un9i46w7m)
* There are a lot of feels in open source. [Geeksta](http://geeksta.net/geeklog/exploring-expressions-emotions-github-commit-messages/) examines how emotions are expressed in GitHub commit messages
* Are bigger pull requests better? Jessie Frazelle analyzed [the top 15 projects on GitHub](https://blog.jessfraz.com/post/analyzing-github-pull-request-data-with-big-query/) in terms of pull requests opened vs. pull requests closed

Visualizations

* [to read] Google [Data Studio 360 dashboard](https://datastudio.google.com/open/0ByGAKP3QmCjLdXBlWVdrZU5yZW8) (previous post [about Data Studio](https://medium.com/google-cloud/showing-off-the-new-free-google-analytics-data-studio-with-reddit-aprils-gilded-comments-for-ebe965dbbb15)).

Outras idéias

* Prever a experiencia pelo estilo de codigo

**Referencias**

* <https://github.com/fhoffa/analyzing_github>
* <https://blog.github.com/2016-06-29-making-open-source-data-more-available/>
* [to read] <https://kozikow.com/2016/06/05/more-advanced-github-code-search/>

“Just as books capture thoughts and ideas, software encodes human knowledge in a machine-readable form.“