Parse the GithubArchive for Fun

What Data it has,
where it comes from,
and how to play with it.

What is GithubArchive

Not an official part of Github

https://github.com/igrigorik/githubarchive.org

https://www.githubarchive.org/

Build from parsing “official” API endpoints for event Data

Event Data

Besides all the usual Metrics and Graphs Github shows, it has an Event Feed to show Timelines in different context.

Examples:

- Push
- Watch
- New Issue
- Pull Request

Get the Data

(copied from website)

Command

Activity for 1/1/2015 @ 3PM UTC

```
wget http://data.githubarchive.org/2015-01-01-15.json.gz
```

Activity for 1/1/2015

```
wget http://data.githubarchive.org/2015-01-01-{0..23}.json.gz
```

Activity for all of January 2015

```
wget http://data.githubarchive.org/2015-01-{01..30}-{0..23}.json.gz
```

Get the Data

Every Line in this File is a self containing json containing one event

Example: https://github.com/community-stats/crawler/blob/master/docs/example_github_event.json {

```
"id":"4343316157",
"type":"PushEvent",
"actor":{
  "id":903479,
  "login":"openstack-gerrit",
  "display_login":"openstack-gerrit",
  "gravatar_id":"",
  "url":"https://api.github.com/users/openstack-gerrit",
  "avatar_url":"https://avatars.githubusercontent.com/u/903479?"
},
"repo":{
  "id":13839311,
  "name":"openstack/openstack",
  "url":"https://api.github.com/repos/openstack/openstack"
},...
```

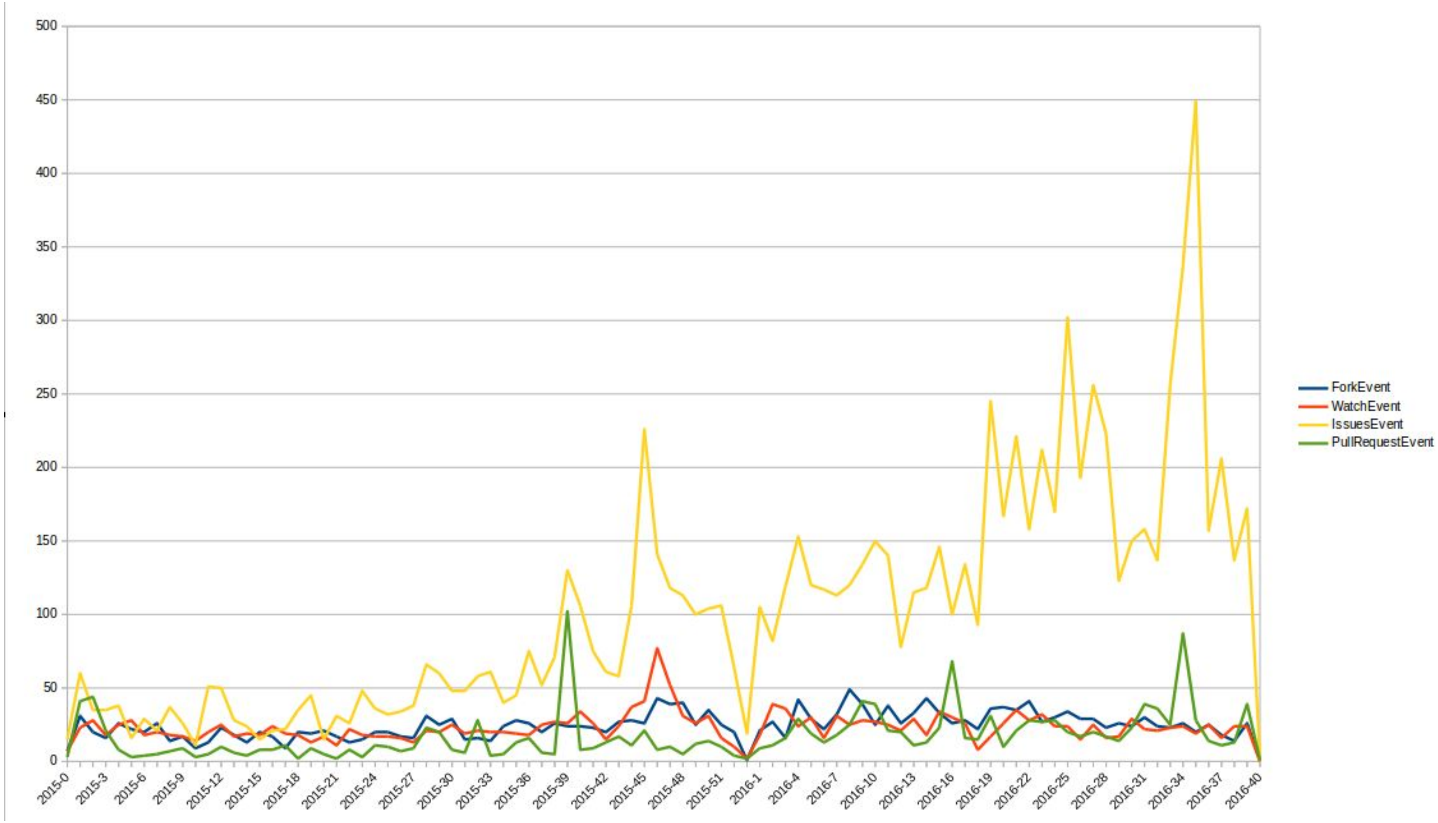
How to parse this big amount of Data

You could use the recommended way of using Googles BigData Query service

Or you build an own Processing Pipeline

- Fetch
  - Filter for wanted entries
  - Only use the needed Values, not the whole entry
  - Add them to your preferred way of storage

Why would I want to?



Why would I want to

- More detailed over Time Interaction Stats
- Combine Stats over different Projects
- Relate Interactions to certain points in Time
  - Releases
  - News Mentions
  - Conferences
  - Meetups

OpenSource projects are **Products**, why not analyze them like one.

Why would I want to

A big Project is not just a single Repository or Github Organization

Its a whole **ecosystem**!

Keep track of all the Plugins, Extensions and Tools belonging to your Project.

Thank You for Listening

Me: Daniel Fahlke aka Flyingmana

Email: flyingmana@googlemail.com

Links again:

- https://github.com/community-stats
- https://www.githubarchive.org/
- https://www.researchgate.net/project/Putting-the-Magento-Open-Source-Community-into-Numbers

Thank *the Berlin PHP Usergroup* for hosting this Meetup