Mining Github and other sources

Overview

This describes the resources that make up the official GitHub REST API v3. If you have any problems or requests, please contact GitHub Support.

i. Current version
ii. Schema
iii. Authentication
iv. Parameters
v. Root endpoint
vi. GraphQL global node IDs
vii. Client errors
viii. HTTP redirects
ix. HTTP verbs
x. Hypermedia

Overview
Media Types
OAuth Authorizations API
Other Authentication Methods
Troubleshooting
API Previews
Versions
Activity
Checks
Gists
Git Data

Rate limit

- For API requests using Basic Authentication or OAuth, you can make up to **5000** requests per hour.

The GHTorrent project

Welcome to the GHTorrent project, an effort to create a scalable, queriable, offline mirror of data offered through the Github REST API.

Follow @ghtorrent on Twitter for project updates and exciting research done with GHTorrent.

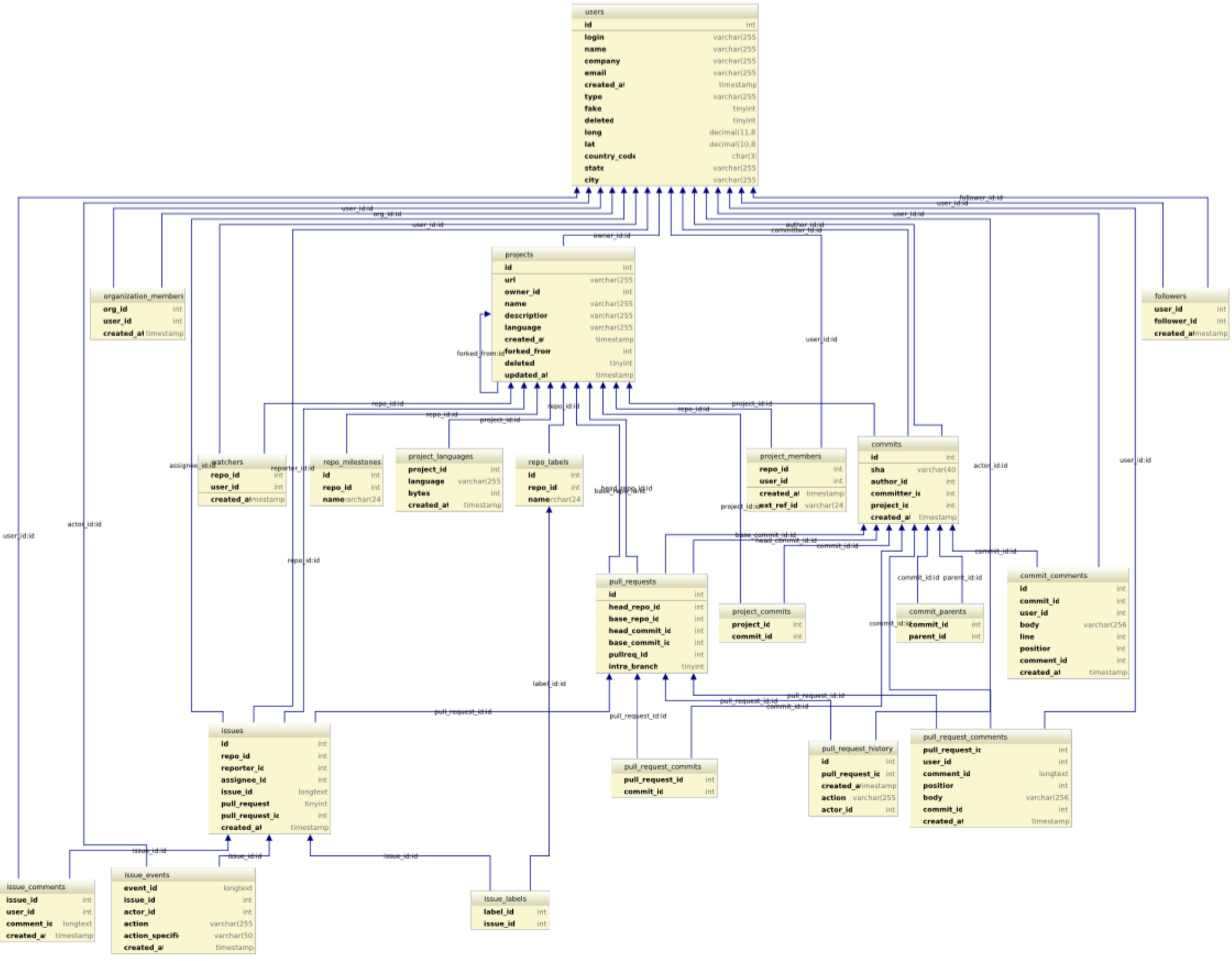
What does GHTorrent do?

GHTorrent monitors the Github public event time line. For each event, it retrieves its contents and their dependencies, exhaustively. It then stores the raw JSON responses to a MongoDB database, while also extracting their structure in a MySQL database.

GHTorrent works in a distributed manner. A RabbitMQ message queue sits between the event mirroring and data retrieval phases, so that both can be run on a cluster of machines. Have a look at this presentation and read this paper if you want to know more. Here is the source code.

The project releases the data collected during that period as downloadable archives.

SQL database



Contains metadata

- Users
- Projects
- Commits
- Pull requests and issues
- Followers
- Watchers…

MongoDb

- Contains the actual contents of the events

GrimoireLab

- By far the best set of tools to mine software repositories
    - https://github.com/chaoss/grimoirelab

Main components

- Perceval: data retrieval
- Sorting hat: identity unification

Perceval

```
    askbot           Fetch questions and answers from Askbot site
    bugzilla         Fetch bugs from a Bugzilla server
    bugzillarest     Fetch bugs from a Bugzilla server (>=5.0) using its REST API
    confluence       Fetch contents from a Confluence server
    discourse        Fetch posts from Discourse site
    dockerhub        Fetch repository data from Docker Hub site
    gerrit           Fetch reviews from a Gerrit server
    git              Fetch commits from Git
    github           Fetch issues, pull requests and repository information from GitHub
    gitlab           Fetch issues, merge requests from GitLab
    googlehits       Fetch hits from Google API
    groupsio         Fetch messages from Groups.io
    hyperkitty       Fetch messages from a HyperKitty archiver
    jenkins          Fetch builds from a Jenkins server
    jira             Fetch issues from JIRA issue tracker
    launchpad        Fetch issues from Launchpad issue tracker
    mattermost       Fetch posts from a Mattermost server
    mbox             Fetch messages from MBox files
    mediawiki        Fetch pages and revisions from a MediaWiki site
    meetup           Fetch events from a Meetup group
    nntp             Fetch articles from a NNTP news group
    phabricator      Fetch tasks from a Phabricator site
    pipermail        Fetch messages from a Pipermail archiver
    redmine          Fetch issues from a Redmine server
    rss              Fetch entries from a RSS feed server
    slack            Fetch messages from a Slack channel
    stackexchange    Fetch questions from StackExchange sites
    supybot          Fetch messages from Supybot log files
    telegram         Fetch messages from the Telegram server
    twitter          Fetch tweets from the Twitter Search API
```

Sorting Hat

- Sorting Hat unifies the identities contributor's identities coming (potentially) from different sources.

https://chaoss.github.io/grimoirelab-tutorial/

Tutorial
Presentation
Introduction
The basics
Perceval
Graal
Producing Kibana dashboards with GrimoireELK
SortingHat: managing identities
Reporting with Manuscripts
SirMordred: orchestrating everything
Python scripting
Cases: CHAOSS Metrics
Tools and tips
Internals
Contributing
License

Table of Contents

- Quick overview
  - Data retrieval
  - Data storage
  - Identities and personal metadata
  - Visualization and analytics

GrimoireLab Tutorial

GrimoireLab is free, open source software for software development analytics. It allows you to retrieve data from many kinds of systems with information related to software development, and produce analysis and visualizations with it.

GrimoireLab supports more than 20 different kinds of data sources, from git repositories or GitHub projects, to issue trackers such as Jira or Bugzilla, including messaging systems such as IRC, Slack or mailing lists, or other types of systems such as StackOverflow or Jenkins.