proof_platform

Scalable Web Scrapping

Installation guide

Tomáš Kocman, Libor Polčák

TARZAN project VI20172020062 document
Faculty of Information Technology, Brno University of Technology

Last change: March 2, 2020

proof_platform — Installation guide

Tomáš Kocman, Libor Polčák

Faculty of Information Technology, Brno University of Technology, e-mail:
polcak@fit.vutbr.cz

This project was created as part of the TARZAN project and the diploma thesis of Tomáš Kocman[1]. This project allows scalable web scrapping, i.e. archiving of web pages.

1 Getting Started

Please instal Docker as the first step[2]. This project is only supported as a Docker image. If you want to run Docker-less, you will need to build `scrapit` and `lemit` (see the subdirectories). For Docker-less install, you will also need PostgreSQL RDBMS[3], Redis[4], pgAdmin[5].

Once you install Docker, edit the `docker-compose.yml volume` section. Select a directory where to store the archives.

Then, installation of the applications is as simple as:

```
docker-compose build
docker-compose up -d
```

[1] https://www.fit.vut.cz/study/thesis/21459/
[2] https://docs.docker.com/install/
[3] https://www.postgresql.org/
[4] https://redis.io/
[5] https://www.pgadmin.org/