INSTALLATION GUIDE

Overview

1. The application is fully dockerized for ease of deployment, nevertheless it is a large system with more than a dozen containers.
2. There are two major subsystems, Sitehound-* and deep-*. You can read more here
3. Ideally these two components are deploying separately and they will communicate via Apache Kafka.
4. For reference, there is an Architecture Diagram shown below.
5. Once deployed you will only interact with the application via your browser.

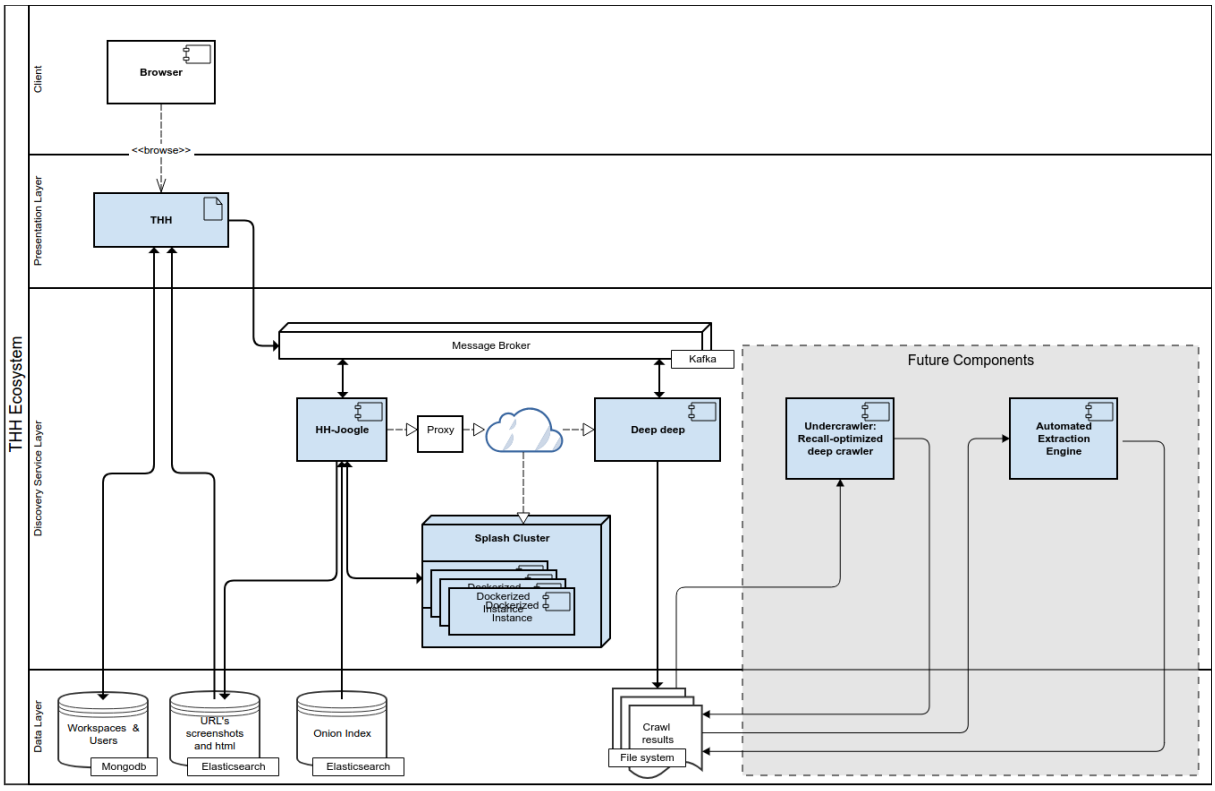

Figure 1: *Architecture diagram*

Provided Hosted version

We are aware that the hardware requirements are not easy to meet so we also provide a hosted version. Send us an email to support@hyperiongray.com and we will set you up and account.

Recommended Hardware

Since the stack of applications contains several infrastructure containers (mongo, elasticsearch, redis) and is designed to take advantage of the multicore architecture, we would recommend:

For a single host:

- At least 100GB of storage (if you plan to do serious crawling, 1TB is better)
- At least 16GB of dedicated RAM, 32GB is better
- 8 CPU cores

Cloud Deployment on Amazon's EC2

1. For a single host installation, we recomend `m4.2xlarge` instance type.
2. On the security groups panel, open the inbound traffic for ports: `5081`, `2181` and `9092` on the Sitehound host EC2.

Prerequisites

1. Ubuntu 16.04 is the recommended OS but it should play well with other distros. It won't work on Windows nor Mac though so a Virtual Machine would be needed in these cases. You can get one here.
2. Update the system:

   ```
   sudo apt update
   ```

3. Docker CE or better installed. Docker API version should be at least 1.24. For Ubuntu 16.04 run:

   ```
   sudo apt install docker.io
   ```

4. docker-compose installed on the Deep-deep server. Version should be at least 1.10. For Ubuntu 16.04 run:

   ```
   sudo apt install -y python-pip
   export LC_ALL="en_US.UTF-8"
   export LC_CTYPE="en_US.UTF-8"
   sudo -H pip install docker-compose
   ```

5. `$USER` (current user) added to the docker group. For Ubuntu 16.04 and user `ubuntu`, run:

```
sudo usermod -aG docker ubuntu
```

and re-login.

Base installation

1. From the provided .zip file, copy the folder ./sitehound-configs to the home directory of the server, or servers if you choose the dual host installation.

   ```
   scp -r sitehound-configs ubuntu@{host-ip}:~
   ```

2. All further actions are executed in this `sitehound-configs` directory:

   ```
   cd sitehound-configs
   ```

Deep-deep installation

Download deep-deep models on the host where the deep-deep will be running:

```
./download-deep-deep-models.sh
```

Running Sitehound + Deep-deep on the same host

1. Make sure port `5081` is open. Start all services with:

```
docker-compose \
    -f docker-compose.yml \
    -f docker-compose.deep-deep.yml \
    up -d
```

Where is the data stored?

Sitehound data is kept in mongodb and elasticsearch databases, outside the containers, using mounted volumes in `volumes/` folder, so if you do `docker-compose down` or remove containers with `docker rm`, data will be persisted.

Crawl results of deep-deep would be in `deep-deep-jobs/`, and broad crawl results will be in `./dd-jobs/`. Crawled data in CDRv3 format will be in `./dd-jobs/*/out/*.jl.gz`, and downloaded media items in `./dd-jobs/*/out/media/`.

What you would be using without installing

The application will be using several external services:

1. Crawlera: a proxy rotator.
2. Our clustered hosted version of Splash.
3. An onion index hosted by SRI.

You are ready!

1. Navigate to http://localhost:5081 (or your Sitehound's IP address).
2. Log in with user `admin@hyperiongray.com` and password `changeme!`.
3. Create a new workspace, and click on the row to select it.
4. Follow this walk-through for a better insight.

Help

You can find a PDF version of this guide here.

You can reach us at support@hyperiongray.com.