Anomaly Detection in CI Jobs

https://etherpad.openstack.org/p/wadci

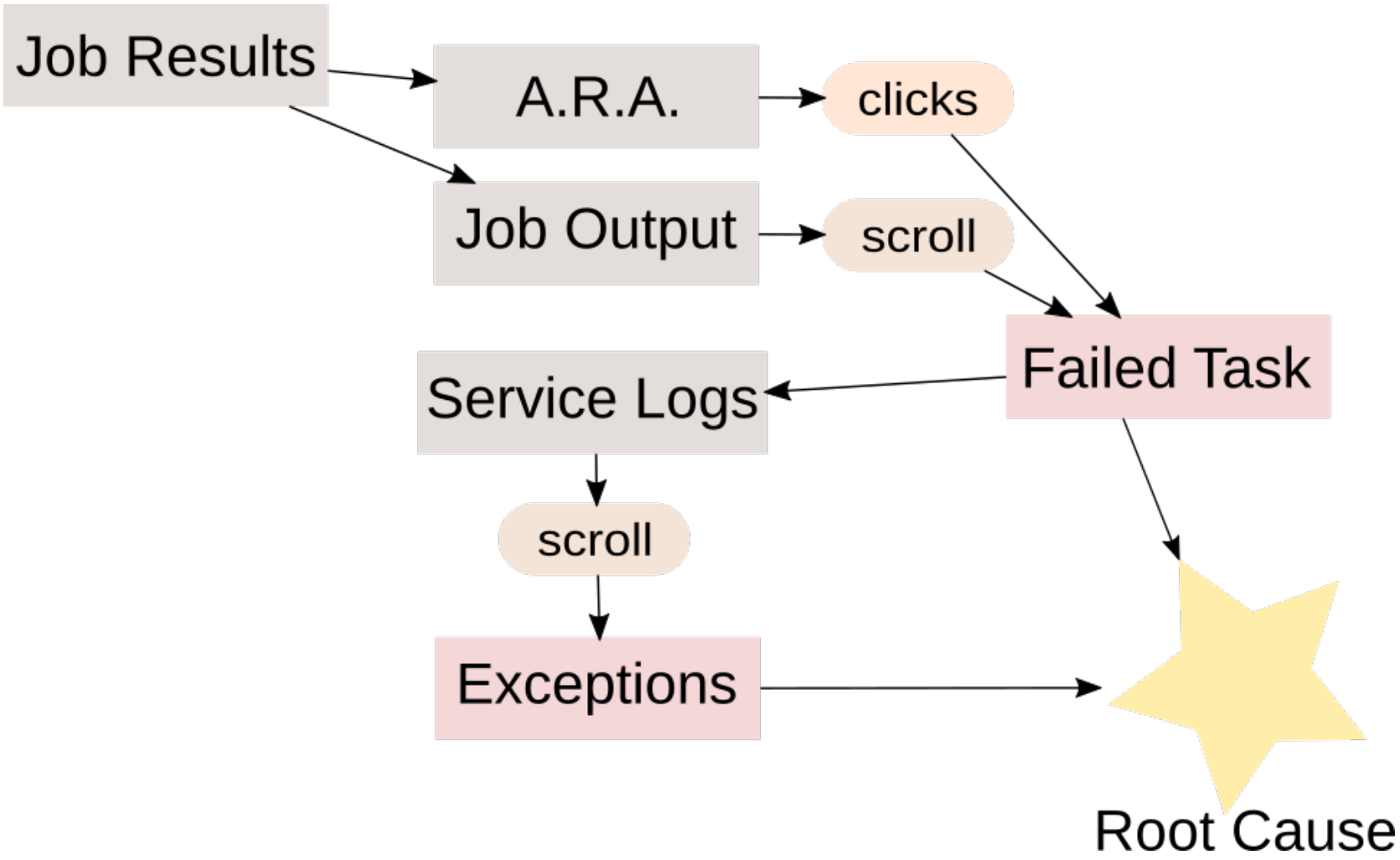
tdecacqu@redhat.com

2018-03-08

Outline

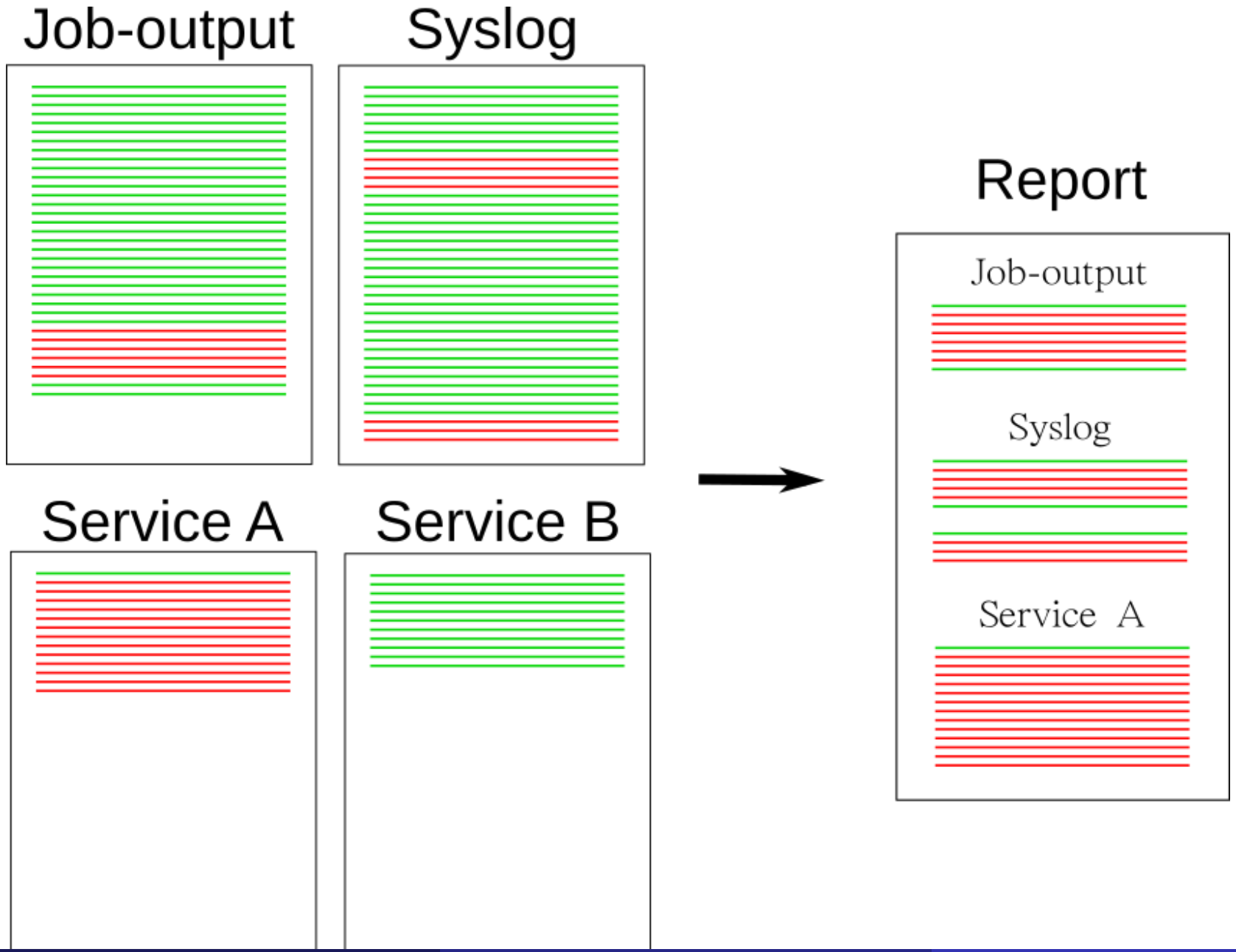
Topic

Current Process



What if the machine looked for the errors?

And produced a nice report?

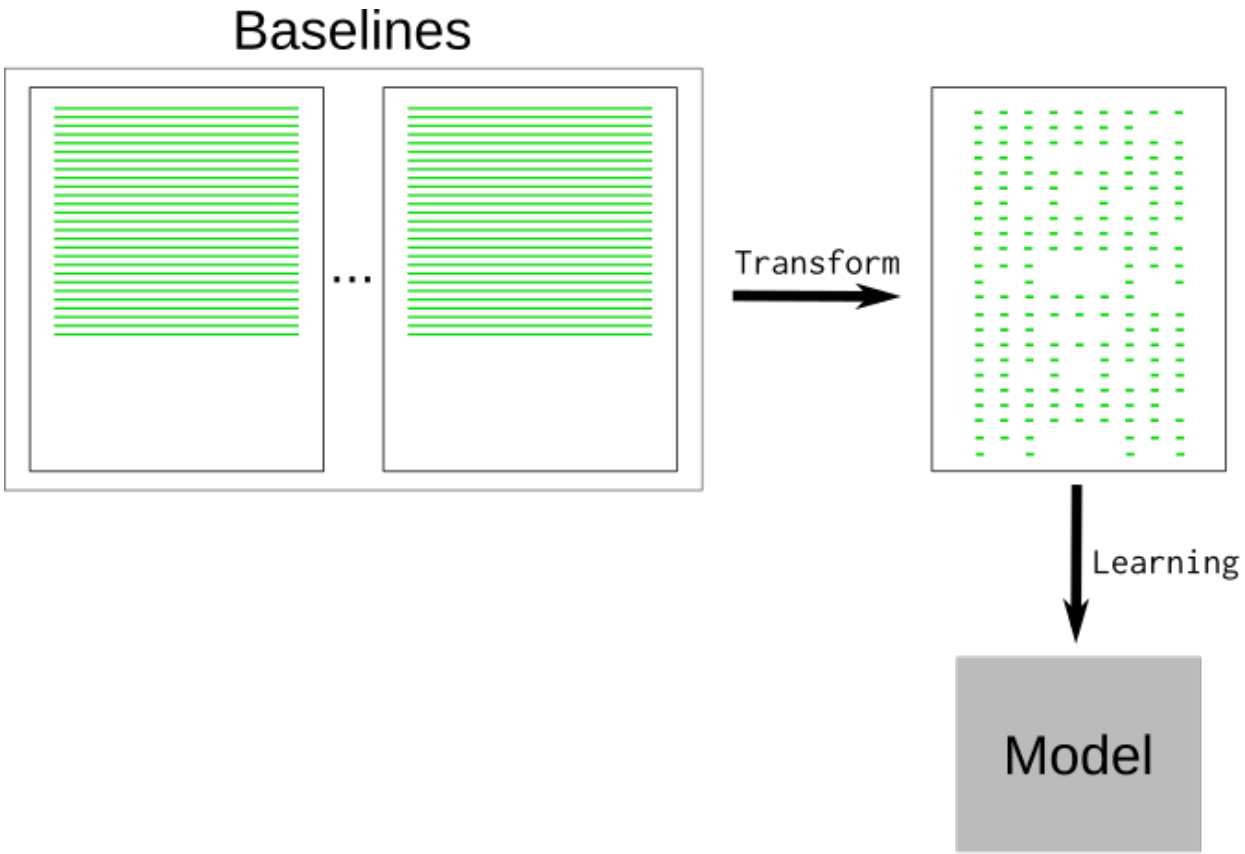


Base Principle

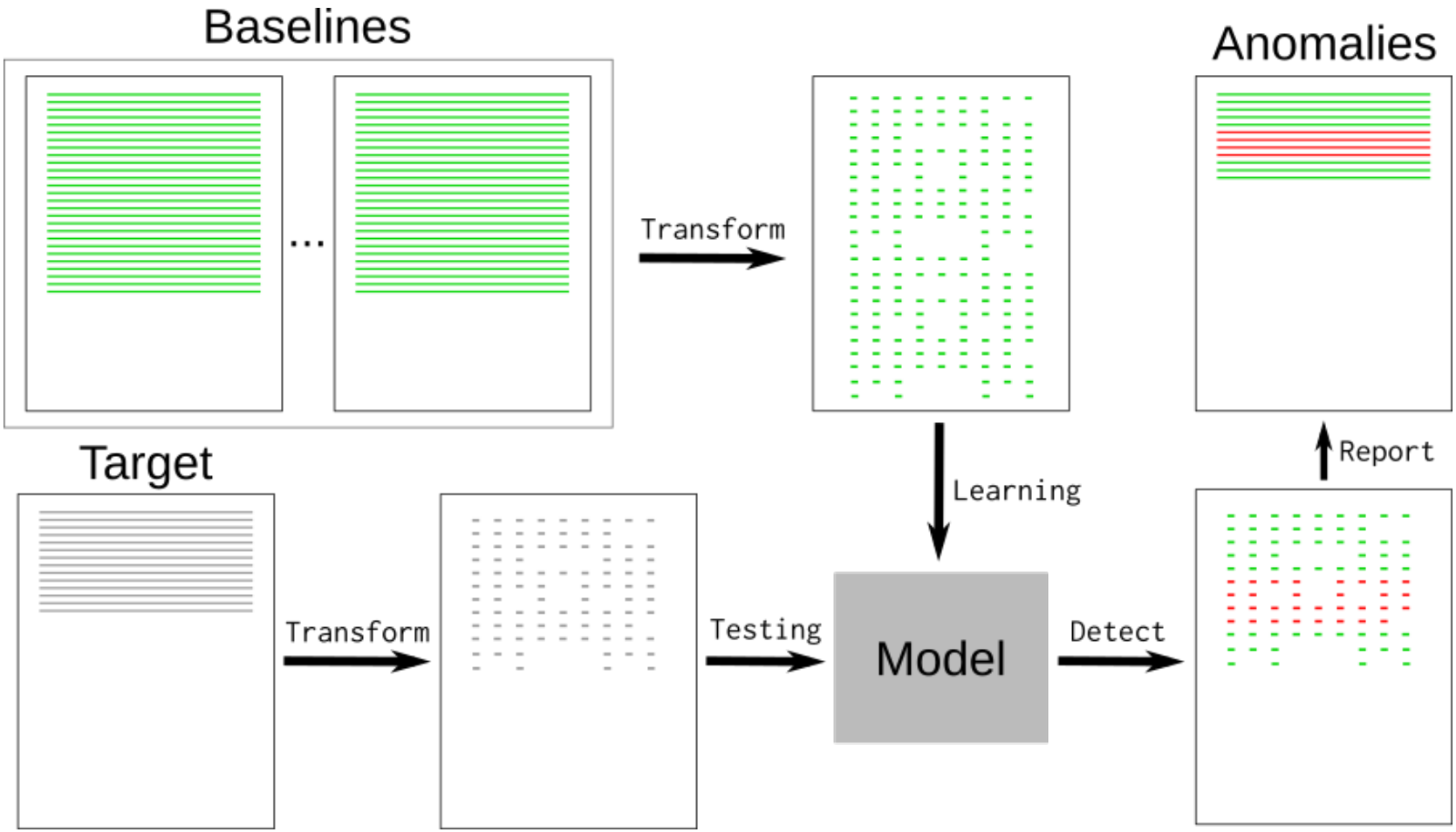
- Baseline: previous job logs
- Target: failed job logs
- Anomaly: new lines missing from the baseline

Topic

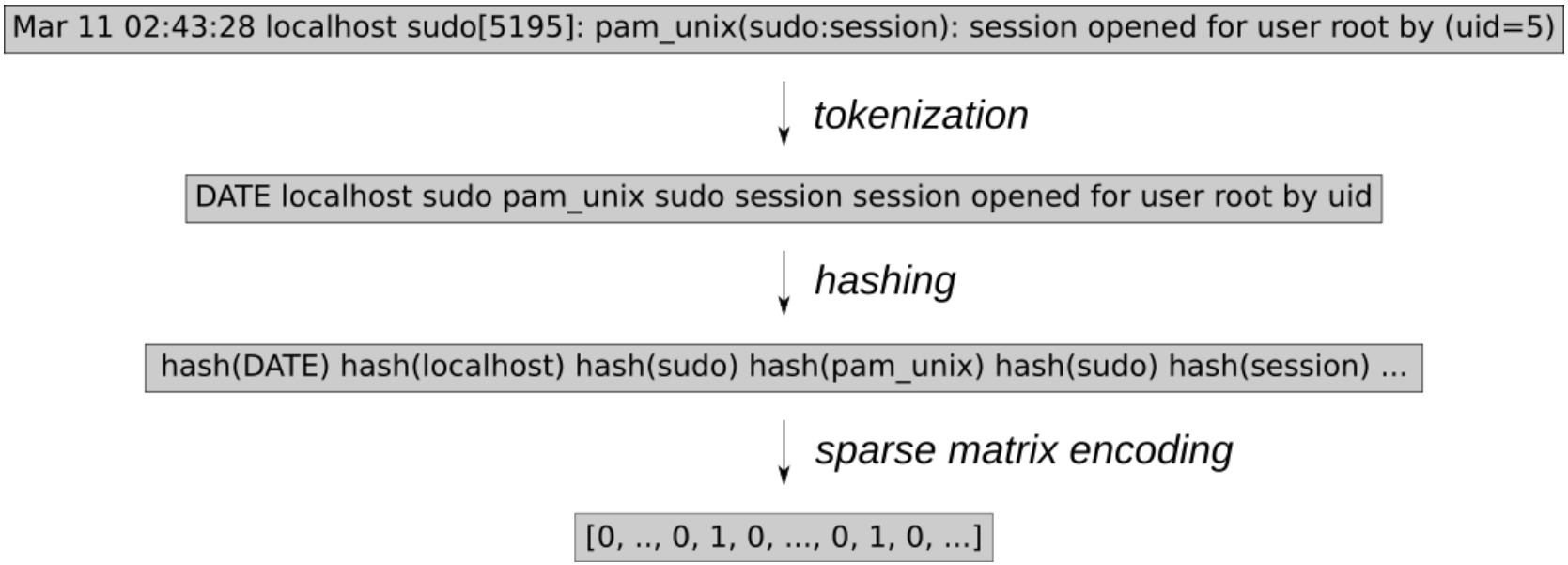
Generic Training Workflow



Generic Testing Workflow



Hashing Vectorizer

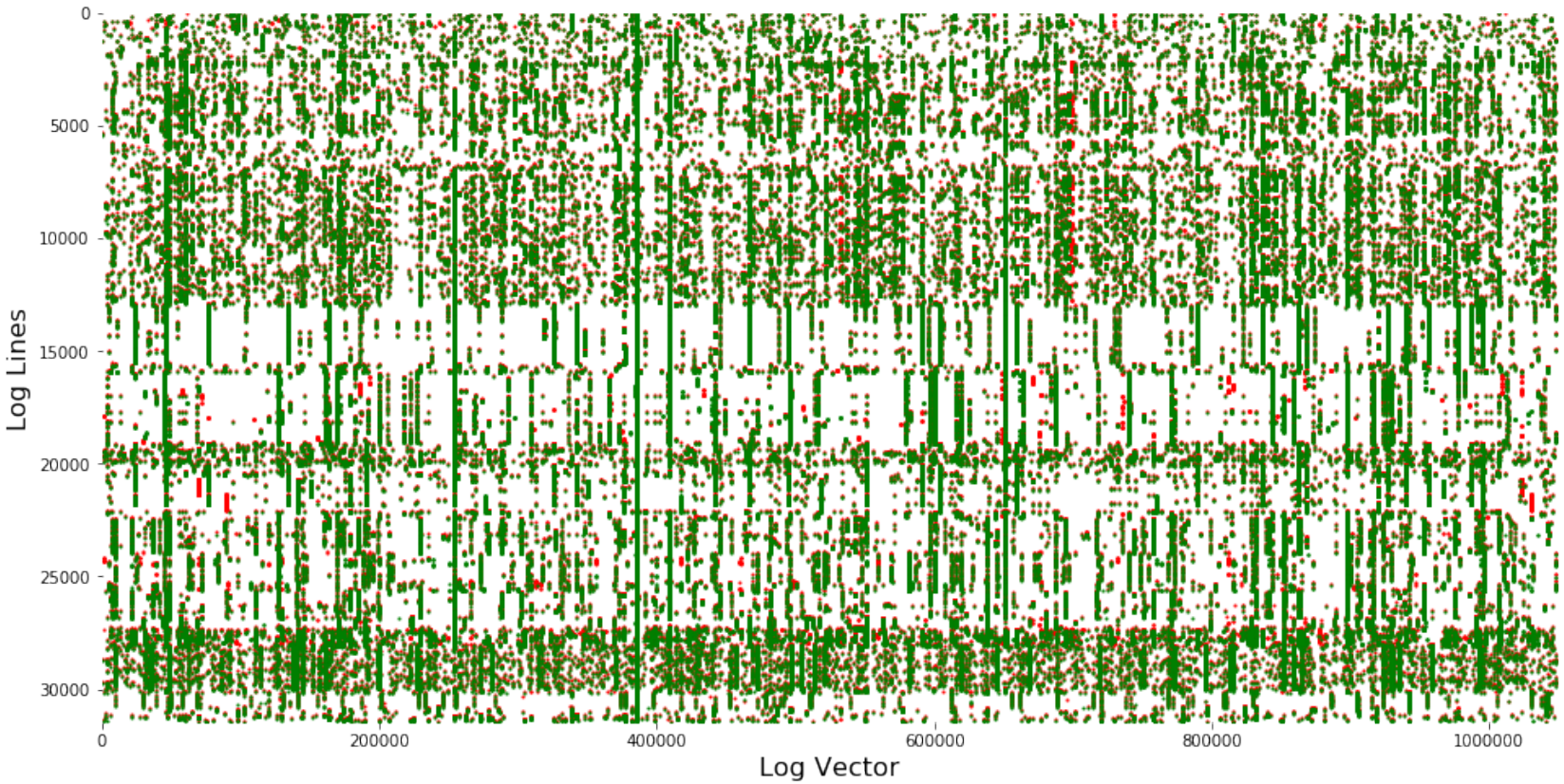


Noise Reduction

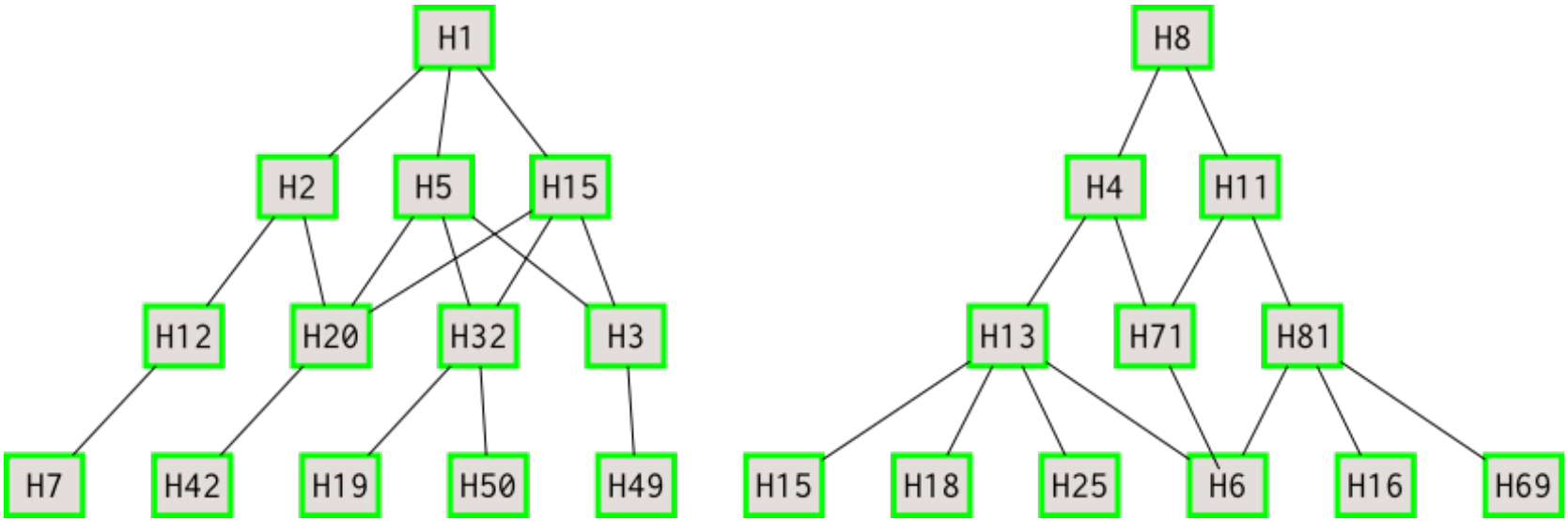
- Random words may be replaced with known tokens:

Token	Raw text
DATE	months/days/date
RNGU	uuid
RNGI	ipv4/ipv6/mac
RNGN	words that are 32, 64 or 128 char long
""	all numbers and non letters

Example of Devstack Vectors



Nearest Neighbors Unsupervised Learner



kNeighbors computes vector's distance

```
2018-02-22 00:18:03.959599 | controller | "ephemeral_device": "VARIABLE IS NOT DEFINED!
```

Vector = `controller` `ephemeral_device` `VARIABLE` `IS` `NOT` `DEFINED`

```
kneighbors(Vector) = 0.9
```

Caveats

- Need DEBUG in baseline logs.

Caveats

- Need DEBUG in baseline logs.
- Noise may hide important information:

```
pcre enabled                | pcre disabled
setup mirror hostnameA      | setup mirror hostnameB
```

Caveats

- Need DEBUG in baseline logs.
- Noise may hide important information:

```
pcre enabled                | pcre disabled
setup mirror hostnameA      | setup mirror hostnameB
```

- Tokenization may need adjustment for small dataset.

Topic

Installation

- Use the container image or install using:

```
sudo dnf install -y python3-scikit-learn python3-aiohttp
sudo dnf install -y python3-pip
pip3 install --user logreduce
```

Compare 2 files

- Output *distance* | *filename:line-number*: anomaly

```
$ pushd 01-files/
$ logreduce diff dib-success.log dib-failure.log
0.250 | dib-failure.log:2258: Package python-setuptools-0.9.8-
                                is obsoleted by python2-setuptoo
```

Compare 2 files

- Output *distance* | *filename:line-number*: anomaly

```
$ pushd 01-files/
$ logreduce diff dib-success.log dib-failure.log
0.250 | dib-failure.log:2258: Package python-setuptools-0.9.8-
                                is obsoleted by python2-setuptoo
```

- Multiple baselines can be used

```
$ logreduce diff audit.log.1 audit.log.2 audit.log
0.614 | audit.log:24516: msg='cwd="/home/centos/logreduce" \
                          cmd="su" terminal=pts/7 res=failed'
```

Compare 2 directories

```
$ pushd 02-dirs/
$ logreduce --debug diff success-*/ failure-*/ \
            --html report.html
INFO  Classifier - Training took 84.141 seconds to ingest 33.4
INFO  Classifier - Testing took 173.464 seconds to test 22.952
99.67% reduction (from 128882 lines to 424)
```

Model Training

- Model can be trained offline first:

```
$ logreduce dir-train sosreport.clf sosreport-good/ other/
INFO Training took 1.696 seconds to ingest 0.513 MB
$ du --si sosreport.clf
66k     sosreport.clf
```

- To be used later:

```
$ logreduce dir-run sosreport.clf sosreport-customer/
0.000 | ansible.log:0012: TASK [Command with long output]
0.626 | ansible.log:0014: fatal: [localhost]: FAILED!
0.364 | syslog:1576: localhost: System clock wrong by 1.417479
99.62% reduction (from 1595 lines to 2)
```

Journald

- Extract novelty from the last day:

```
$ logreduce journal --range day
```

- Build a model using last month's logs and look for novelties in the last week:

```
$ logreduce journal-train --range month journald.clf
$ logreduce journal-run   --range week  journald.clf
```

Zuul Jobs

- Build a model

```
$ logreduce job-train model.clf
    --job devstack
    --include-path logs/
    --pipeline gate
    --project openstack-dev/devstack
    --zuul-web http://zuul.openstack.org/api
```

- Re-use the model

```
$ logreduce job-run model.clf http://logs.openstack.org/...
```

Zuul Jobs

- Build a model

```
$ logreduce job-train model.clf
    --job devstack
    --include-path logs/
    --pipeline gate
    --project openstack-dev/devstack
    --zuul-web http://zuul.openstack.org/api
```

- Re-use the model

```
$ logreduce job-run model.clf http://logs.openstack.org/...
```

- Extract anomalies from a job result:

```
$ logreduce job http://logs.openstack.org/...
```

Zuul Jobs Example: tempest-full

- Model trained with:

```
$ logreduce job-train tempest.clf
      --job tempest-full
      --include-path controller/
      --count 3
      --zuul-web http://zuul.openstack.org/api
```

- Usage:

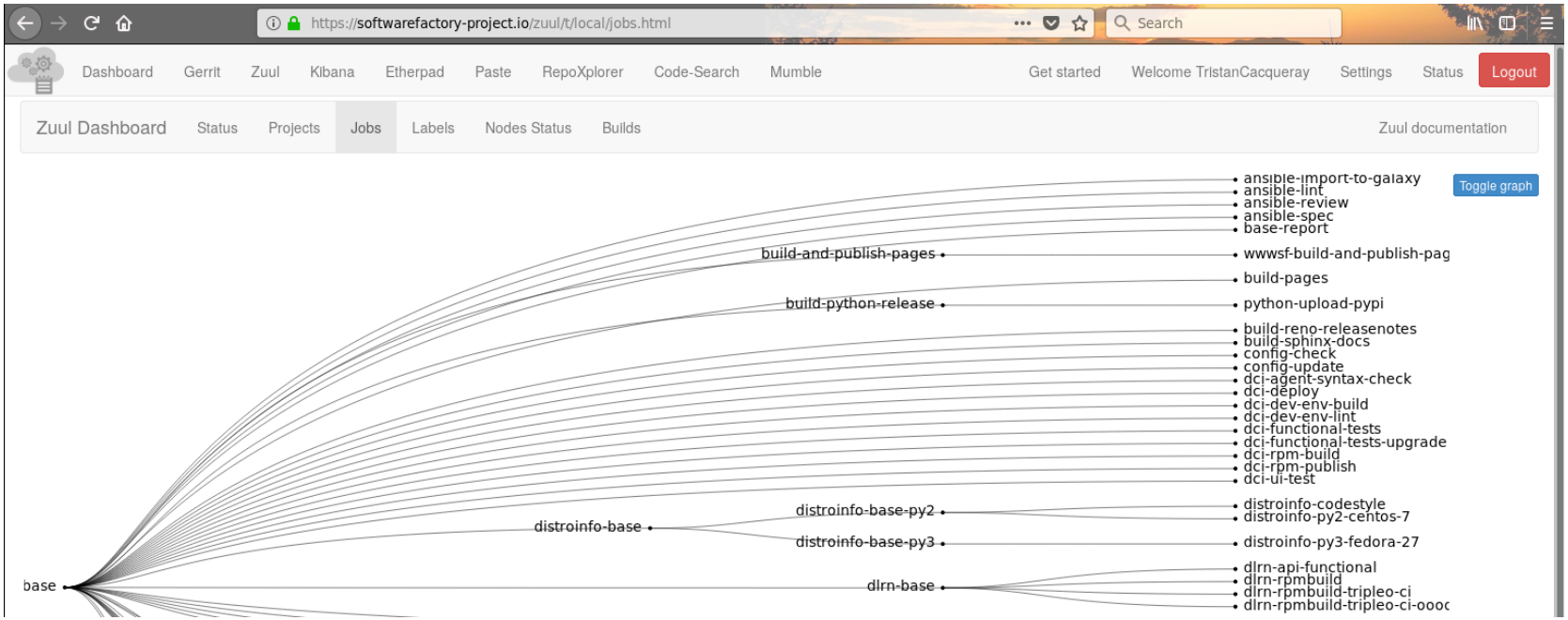
```
$ pushd 03-jobs/
$ logreduce job-run _models/tempest.clf $log_url
      --include-path controller/
```

Command line interface summary

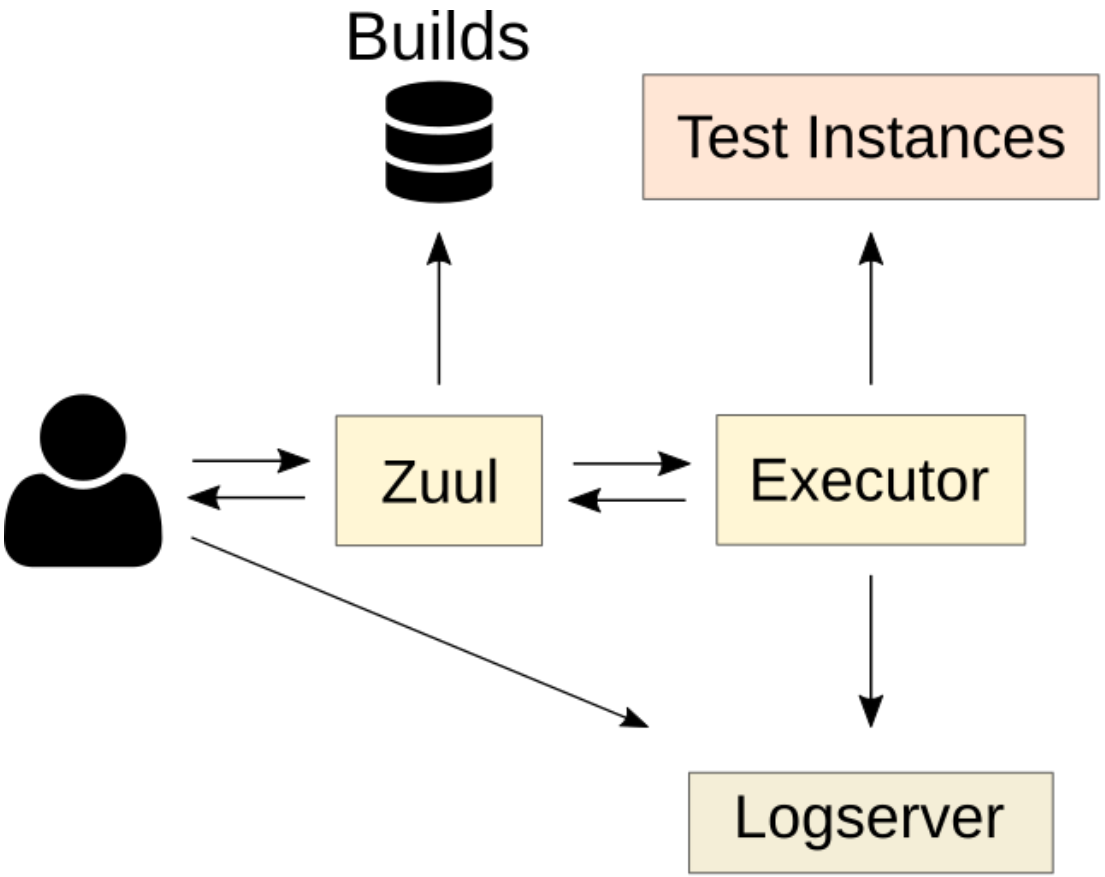
- Supports directories, journald and Zuul jobs.
- Model can be trained *dir-train*, *jounal-train* and *job-train*.
- Model can be re-used: *dir-run*, *journal-run* and *job-run*.
- Or all in one command: *dir*, *journal* and *job*.

Topic

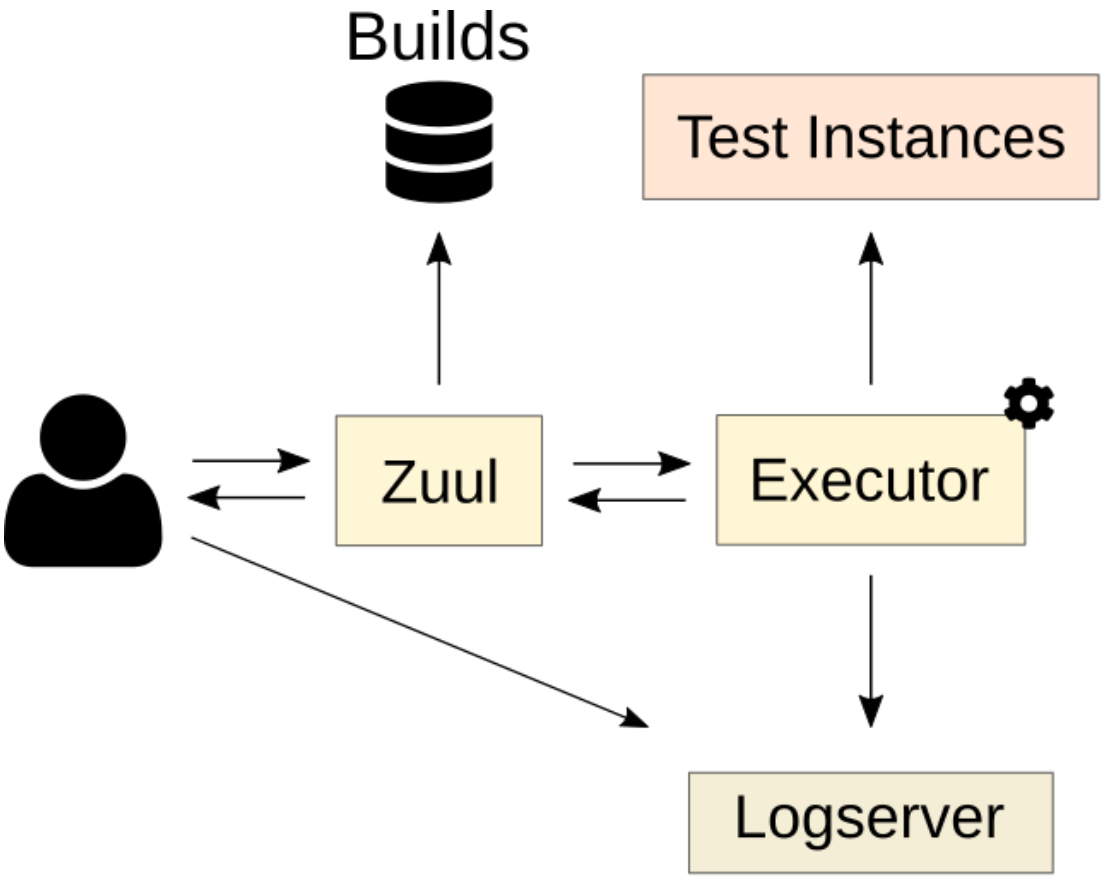
www.softwarefactory-project.io



Zuul Architecture



Post-Run Analysis

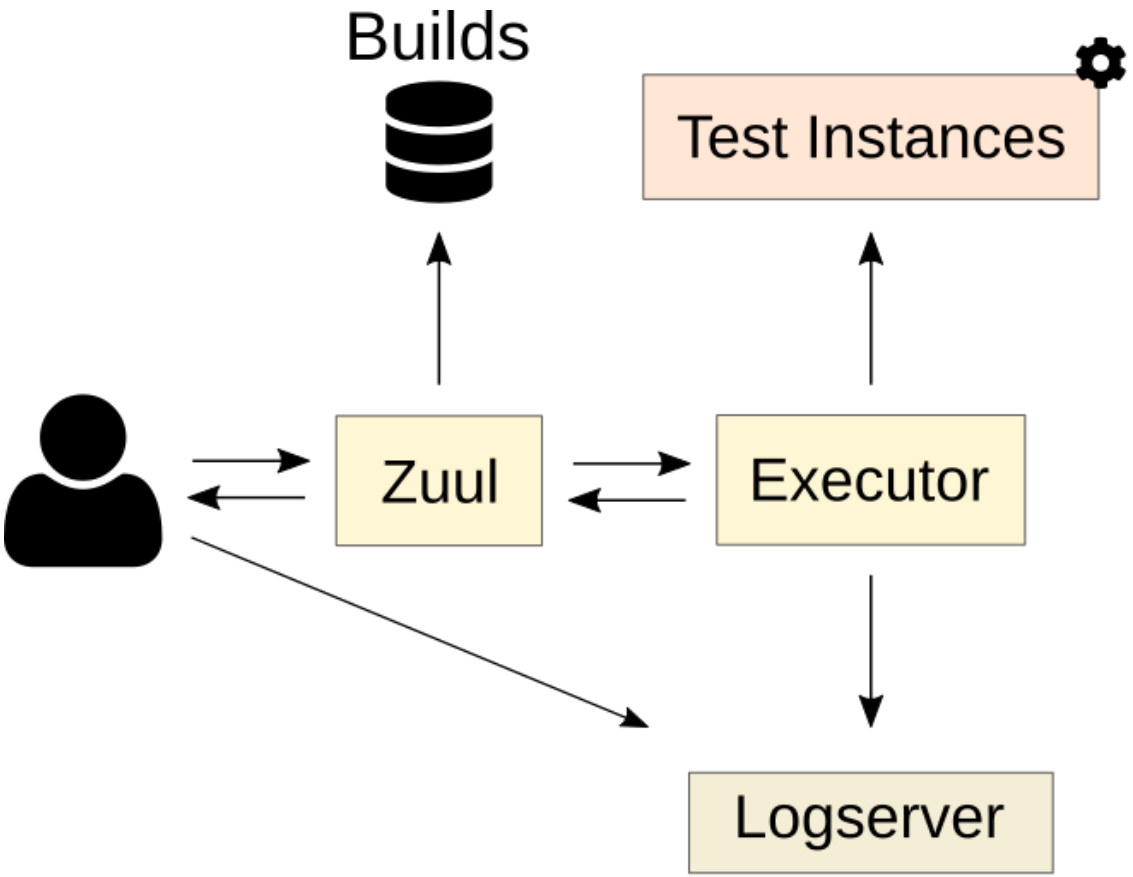


Post-Run Playbook

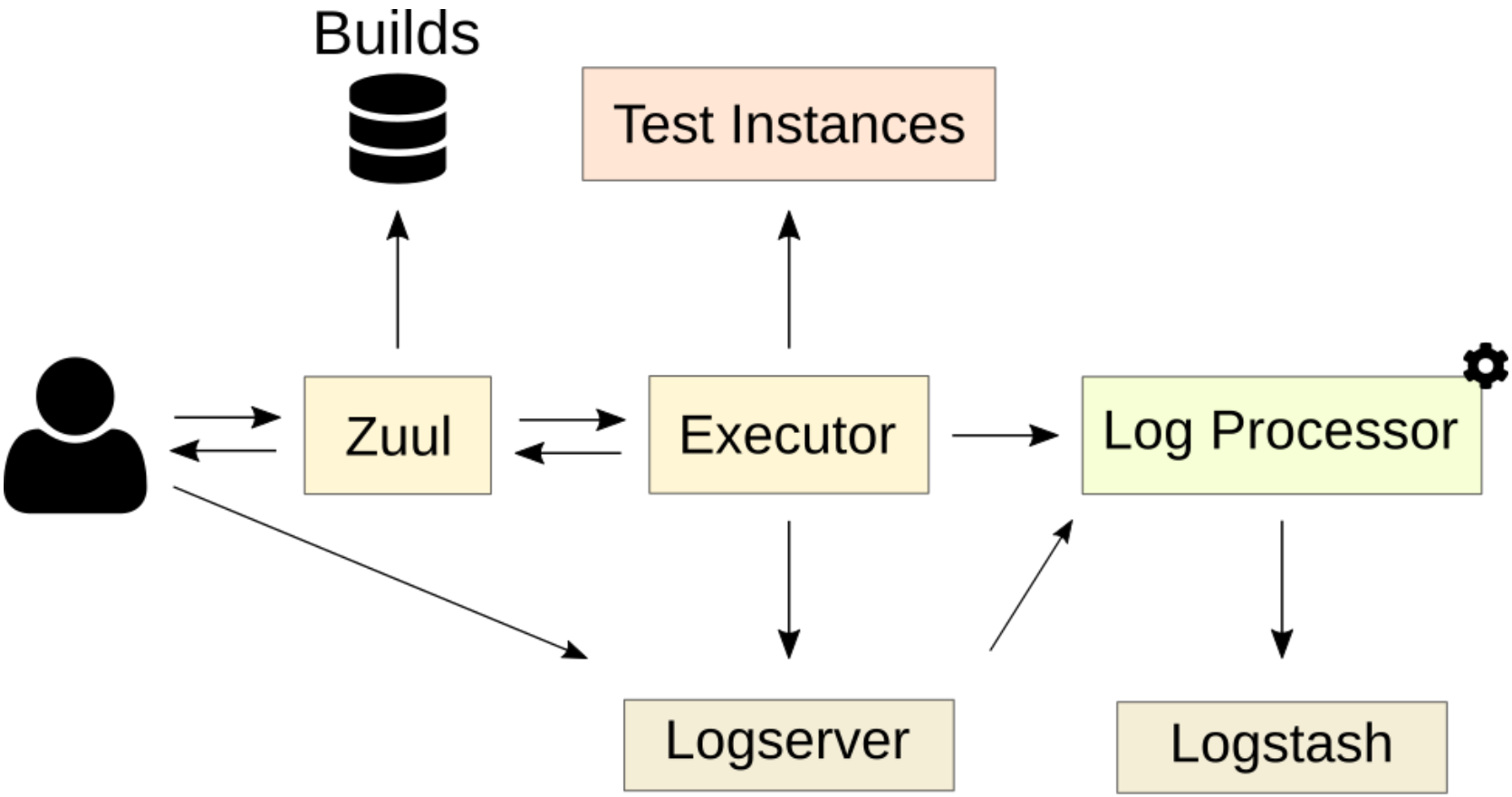
```
- job:
    name: base
    post-run:
      - upload-log
      - clasify-log

- tasks:
  - name: Fetch or build the model
    command: log-classify job-build ...
  - name: Generate report
    command: log-classify job-run ...
  - name: Return report url
    zuul_return: {zuul: url: log: ...}
  - name: Upload model
    synchronize: ...
```

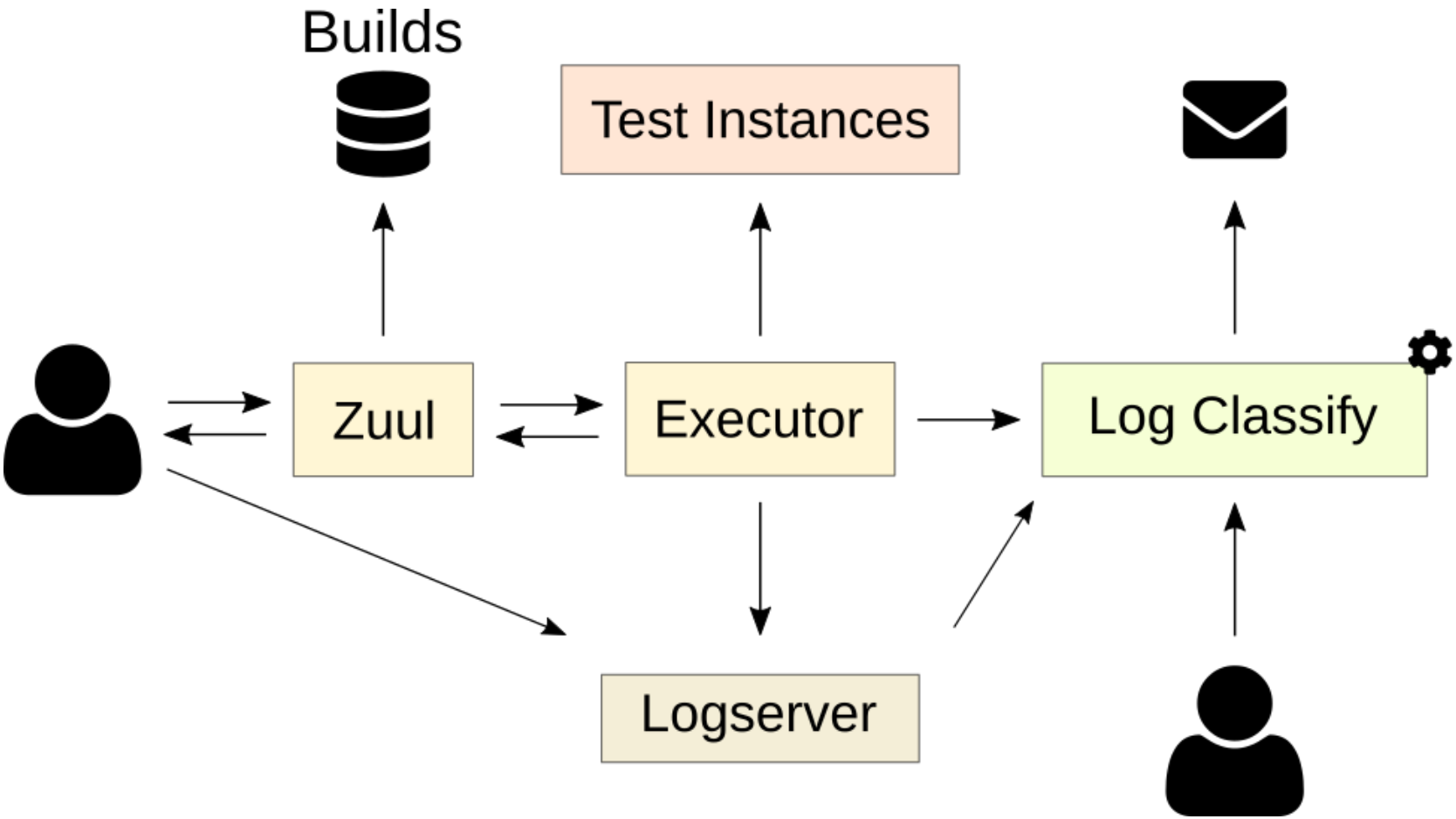
Post-Run Analysis running on test instances



Logstash Filter



Standalone Service



Topic

Credits

- Roadmap:
  - Bootstrap community project.

Credits

- Roadmap:
  - Bootstrap community project.
  - Better supports more jobs.
  - Interface with elastic-recheck.
  - Integrate in openstack-infra.

Credits

- Roadmap:
  - Bootstrap community project.
  - Better supports more jobs.
  - Interface with elastic-recheck.
  - Integrate in openstack-infra.

- Icons used in diagrams are licensed under Creative Commons Attribution 3.0: `https://fontawesome.com/license`