invenio-stats

Nicolas Harraudeau
Dinos Kousidis

Agenda

- Requirements
- Architecture
- Current state
- Further development

Requirements

Publisher needs to know how many times his files have been downloaded

People in charge of communities need to know which records are most viewed

Who are the top uploaders

Ratio between open access and closed access records

What it provides

Measures the usage of records and the activity of users and communities

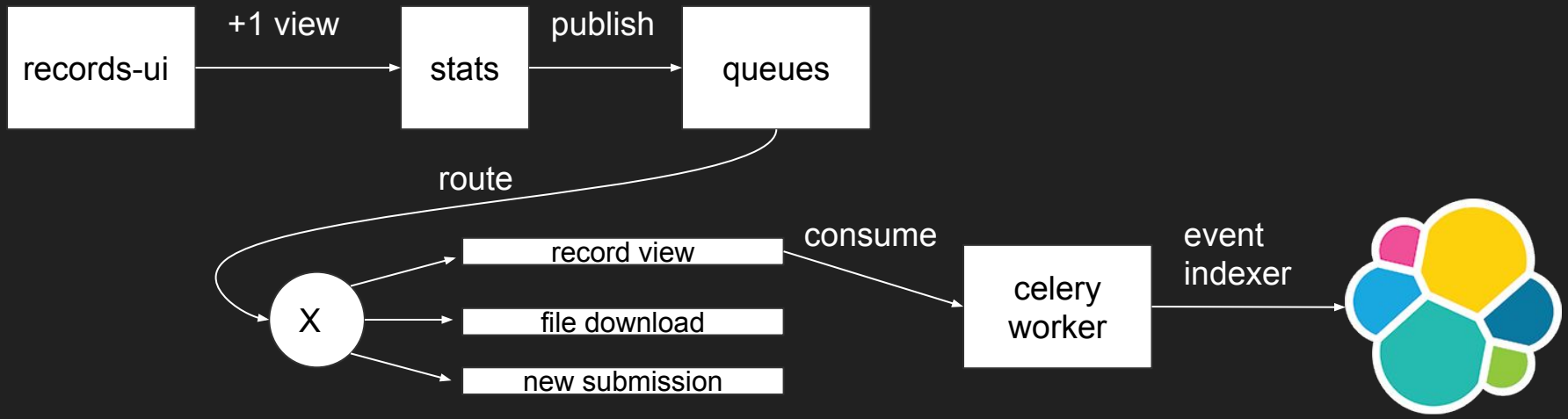
Data required for the calculations acquired by monitoring low level events or exist already in DB

Statistics are stored in ES

Architecture

1. Events are monitored using signals
   - invenio-records-ui → Record views
   - invenio-files-rest → File downloads
2. Events are passed to invenio-queues
3. Collected by a celery worker to get indexed in ES

Procedure of registering an event



Record view event in detail

```
current_stats.publish('record_view', [dict(
    # When:
    timestamp=datetime.utcnow().isoformat(),
    # What:
    id=record.id,
    pid_type=pid.pid_type,
    pid_value=pid.pid_value,
    labels=record.get('communities', []),
    # Who:
    **get_user()
)])
```

invenio-queues

Based on previous work in invenio-indexer

Decouples the choice of message queue broker

Provides an API for the queue actions e.g. publishing and consuming

Any module that uses a queue can declare it via the 'invenio_queues.queues' entrypoint and access it from the invenio-queues proxy

Elasticsearch as primary data store

ES is not advertised as such but many use it for this purpose

Resiliency has improved a lot from earlier versions

Audit log is most critical but has a low priority

Elastic Cloud provides snapshots every 30 minutes if needed

Some downtime is okay in case of failure

RabbitMQ can also guard from data loss

Elasticsearch

2 types of indices

Short term: 1 index per day per event: event-file-download-19-06-2017

Long term: indices for aggregated statistics

Templates placed in ES with `invenio index init` by invenio-search

Further Development

Add events

Write celery tasks

Solve challenges on how to measure statistics properly

COUNTER provides general guidelines

Making stats COUNTER compliant

A standard on how to count statistics of online library data defining rules on:

- Double clicks (refresh, back button) 10s
- File downloads 30s
- Bot searches

Defines format to produce monthly reports

Links

Requirements: https://github.com/inveniosoftware/invenio-stats/wiki/Requirements

CNIL: https://www.cnil.fr/sites/default/files/atoms/files/cnil_en_bref-ven-2017-vd.pdf

ES as primary data storage: https://discuss.elastic.co/t/elasticsearch-2-3-as-primary-data-store/50265/3

ES snapshots: https://www.elastic.co/guide/en/elasticsearch/reference/1.7/modules-snapshots.html

COUNTER: http://irus.mimas.ac.uk/documents/IRUS-UK_COUNTER_OR2016.pdf

https://www.projectcounter.org/wp-content/uploads/2016/03/Technical-pdf.pdf